JURNAL INFORMATIKA
Volume 13 No. 1 April 2026, pp. 37-48
ISSN (Online) 2528-2247 | ISSN (Print) 2355-6579
DOI: https://doi.org/10.31294/ji.v13i1.11512

Semi-Supervised Bullying Detection in Narrative Student Counselling Reports Using a Hybrid CNN-LSTM with Pseudo-Labelling

Suwarno [1*], Muthia Andini[1], Mangapul Siahaan[1]

[1]Universitas Internasional Batam
Baloi-Sei Ladi, Jl. Gajah Mada, Tiban Indah, Kec. Sekupang, Kota Batam, Indonesia

Correspondence e-mail: suwarno.liang@uib.ac.id

Submission:	Revision:	Acceptance:	Available Online:
30-12-2025	15-01-2026	05-02-2026	12-02-2026

Abstract - Bullying incidents in schools are often documented in narrative student counselling reports containing informal language, emotional expressions, and contextual dependencies, which pose challenges for automated text classification, particularly under limited labeled data conditions. This study aims to develop a bullying detection model for narrative student counselling reports using a Hybrid CNN-LSTM architecture combined with a pseudo-labelling-based semi-supervised learning approach. The proposed model is trained through a two-stage process, consisting of pre-training on approximately 70,000 publicly available abusive-language texts and fine-tuning using 1,000 anonymized student counselling reports validated by guidance counsellors. Pseudo-labelling is employed to expand the training data while preserving domain relevance and adhering to ethical considerations. Experimental results show that the proposed model achieves an accuracy of 0.8698, a recall of 0.8570, and an F1-score of 0.7951. Although the precision value (0.7415) is relatively lower, higher recall is prioritized to reduce the risk of overlooking potential bullying cases in the school counselling context. Comparative analysis with Logistic Regression and Linear SVM indicates that the Hybrid CNN-LSTM model demonstrates more stable performance when processing longer narrative inputs that require contextual interpretation. This study contributes empirical evidence on the effectiveness of semi-supervised deep learning for bullying detection in low-resource, narrative student counselling data, a setting that remains underexplored in prior work.

Keywords: Bullying detection, Deep learning, Hybrid CNN-LSTM, Student counselling reports, Natural language processing

1. Introduction

Bullying remains a persistent problem in Indonesian schools and continues to pose serious risks to students' psychological well-being. National reports indicate a sharp increase in school bullying cases, particularly peer-to-peer incidents, while survey data show that a substantial proportion of Indonesian adolescents experience repeated bullying (KPAI, 2023; UNICEF, 2021). Similar patterns are also observed at the local level, suggesting that bullying occurs across both school and online environments (Hamapu, 2024; Yuliandra, 2025). Despite its high prevalence, bullying often goes unreported due to stigma, fear of disclosure, and limited access to reliable reporting mechanisms, hindering early prevention efforts and increasing the risk of long-term psychological harm (Alhakim et al., 2022).

Traditional counselling systems rely heavily on manual documentation and subjective assessment, resulting in unstructured narrative archives that are difficult to analyze systematically and may delay timely intervention. In response to these limitations, recent advances in Natural Language Processing (NLP) offer promising tools for automating the analysis of counselling reports, particularly through deep learning models capable of capturing emotional cues, abusive language, and implicit indicators of bullying (Purba et al., 2024). However, most existing bullying detection studies are trained on short and informal social media texts, limiting their applicability to longer and emotionally expressive student counselling reports commonly found in educational settings (Akar, 2024; Setiawan et al., 2022).

Another major challenge in educational NLP applications is the scarcity of labeled counselling data, as authentic student reports are sensitive and confidential. Consequently, most prior approaches rely on fully supervised learning paradigms, which are difficult to implement in real-world school environments due to ethical, privacy, and data availability constraints. As a result, empirical evidence on the effectiveness of semi-supervised deep learning models for bullying detection in low-resource counselling settings remains limited, particularly for long narrative texts requiring deeper contextual and emotional understanding.

To address these challenges, recent studies have explored semi-supervised strategies such as pseudo-

Copyright © 2026 Authors
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

labelling for domain adaptation under limited labeled data conditions (Rahamim et al., 2022). In parallel, Hybrid CNN-LSTM architectures have demonstrated strong capability in capturing both local lexical patterns and long-range contextual dependencies, making them suitable for analysing narrative texts (Ullah et al., 2024). Nevertheless, the combined application of pseudo-labelling and Hybrid CNN-LSTM architectures for bullying detection in student counselling reports remains underexplored.

Accordingly, this study addresses two central research questions. First, it examines the effectiveness of a Hybrid CNN-LSTM architecture in detecting bullying within long and narrative student counselling reports. Second, it investigates whether a pseudo-labelling-based semi-supervised learning strategy can enhance bullying detection performance under conditions of limited labeled counselling data. Through this investigation, this study contributes empirical evidence on the applicability of semi-supervised deep learning models for bullying detection in low-resource, real-world school counselling environments.

2. Research Methods

Figure 1 illustrates the research workflow of the proposed bullying detection system. This study adopts an experimental, model-driven approach focusing on the design, implementation, and evaluation of a Hybrid CNN-LSTM architecture for anonymous text-based bullying detection (Hafiza & Setiawan, 2025).

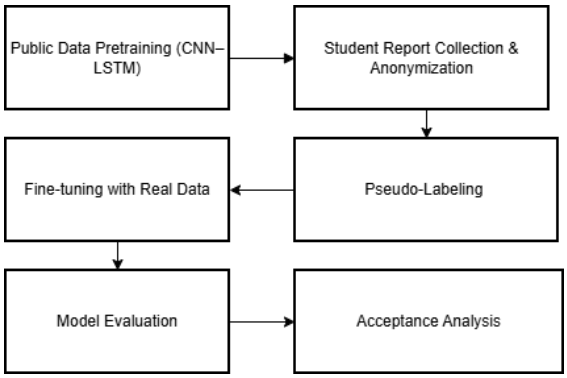

Figure 1. Research workflow of the proposed bullying detection system.

The training process is conducted in two stages: pre-training on large-scale publicly available datasets and fine-tuning on real student bullying reports. This two-stage strategy supports effective domain adaptation in low-resource educational settings, where access to labeled data is constrained by ethical and privacy considerations (Le et al., 2019). An ablation study was not conducted due to data sensitivity and annotation constraints associated with student counselling reports. Nevertheless, the architectural design of the Hybrid CNN-LSTM model was informed by established findings in prior bullying detection and text classification studies.

2.1. Datasets

Two datasets are used in this study: a large-scale public dataset for pretraining and a real-world student report dataset for fine-tuning. The pretraining dataset consists of over 70,000 publicly available social media comments labeled as hate speech, abusive language, or cyberbullying, sourced from multiple platforms and published on GitHub (Andini, 2025). The dataset contains informal, colloquial, and code-mixed Indonesian text, enabling the model to learn general abusive-language patterns under low-resource conditions (Walsh et al., 2021).

The fine-tuning dataset comprises anonymized student reports collected through a web-based anonymous reporting system, describing social-emotional conditions, peer interactions, and suspected bullying incidents. The reports exhibit noisy user-generated characteristics, including non-standard spelling, informal language, code-mixing, and emotionally expressive narratives. Prior studies indicate that hybrid deep-learning models are effective for handling such linguistic variability in Indonesian texts (Y. Zhang et al., 2023).

To address the limited availability of labeled student reports, a pseudo-labelling strategy was applied, where a pretrained model generated initial labels for unlabeled data, followed by expert validation. This semi-supervised approach reduces annotation costs while improving domain adaptation in low-resource NLP settings (Hedderich et al., 2021; Rahamim et al., 2022). To prevent data leakage, overlap and duplication filtering was conducted using TF-IDF cosine similarity. Text pairs with similarity scores above 0.90 were removed, along with exact duplicates. The results confirmed that no overlapping instances existed between the public datasets and the real student report dataset, ensuring a strict separation between pretraining and fine-tuning data. Regarding dataset composition, the total dataset consists of 82,694 text instances. Approximately 70,000 instances originate from publicly available datasets and were used exclusively during the pretraining stage. The real student report dataset contains 1,000 instances collected from school environments and was used solely for fine-tuning and evaluation. The class

distribution of the real dataset is reported separately to ensure transparency and to avoid masking class imbalance within the much larger public dataset (X. Yang et al., 2023; Yu & Zhao, 2021).

2.1.1. Public Dataset

The public dataset used for pretraining was collected from multiple open cyberbullying and aggressive-language corpora. Although the original sources contain various metadata attributes, this study only utilizes the textual content and the corresponding binary labels (bullying = 1, non-bullying = 0). Other attributes were excluded to ensure consistency across sources and compatibility with the student counselling dataset used in the fine-tuning stage. The characteristics and examples of the public dataset used in the pretraining stage are summarized in Table 1.

Table 1. Examples from the Public Bullying Dataset

Label	Text
1	makannya segentong buset Eng. ver: You eat like a whole pot, seriously.
1	mirip pait di little mermaid Eng. ver: You look like Ursula from The Little Mermaid.
0	mu gemes banget adminnya aku ga jadi marah Eng. ver: You're so adorable, admin. I'm not angry anymore.
0	Siang Happy Friday sholat jum'at nya jangan lupa ya aku habis ada kegiatan terus nanti ada pait bareng Team JS Eng. ver: Good afternoon, Happy Friday. Don't forget Friday prayers. I just finished an activity and later there will be practice with Team JS.
0	Libya casualty report … Your bias is showing asswipe cover it up before you stain Wikipedias name even more Eng. ver: Libya casualty report… Your bias is showing, cover it up before you damage Wikipedia's name even more.
0	Just for the record that IP is blocked for hours already Eng. ver: Just for the record, that IP has already been blocked for hours.
1	Stop Editing The Cannistraro Page Just Stop Ok Nothing You Do Is Helping You Are An Idiot Stop Eng. Ver: Stop Editing The Cannistraro Page. Just Stop. Nothing You Do Is Helping. You Are An Idiot. Stop.
1	Fucking whore… You are such a dumb bitch I I make any changes whore Eng. ver: You're a stupid person. I didn't make any changes.
0	"makan nasi padang aja begini badannya" Eng. ver: "Just eating Nasi Padang already makes the body like this."
1	"Makin jelek aja anaknya, padahal ibu ayahnya cakep2" Eng. ver: "The kid keeps getting uglier, even though the parents are good-looking."
1	untuk rakyat atau untuk mereka? Eng. ver: Is this for the people or for them?
0	melindungi pekerja! Eng. ver: Protecting workers!
1	yang kayak gini nih bikin buat childfree aja Eng. ver: Things like this make people choose to be childfree.
0	ngeliat park shin hye abis melahirkan "pait childfree" Eng. ver: Seeing Park Shin Hye after giving birth makes you wonder, "What is childfree?"

2.1.2. Real Student Report

Although Table 2 only presents four representative examples, the real counselling-report dataset used in this study consists of 1,000 anonymized student reports collected from three secondary schools. The samples displayed in Table 2 are provided solely for illustration purposes due to privacy restrictions and publication space limitations. The complete dataset is fully utilized throughout all experimental stages, including preprocessing, labelling, pseudo-labelling, validation, and fine-tuning.

As shown in Table 2, the reports contain key fields such as problem category, narrative description, follow-up status, counselor remarks, and severity level, which together provide sufficient contextual information for the bullying detection task. Since only a subset of the reports initially contained expert-assigned labels, a pseudo-labelling strategy was applied, followed by human validation to ensure label reliability. Only validated samples were included in the final dataset used for model fine-tuning.

Table 2. Representative Examples from the Student Counselling Report Dataset

kategori_masalah Eng. ver: Problem Category	deskripsi_laporan Eng. ver: Report Description	status_tindak_lanjut Eng. ver: Follow-up Status	catatan_guru Eng. ver: Teacher's Notes	tingkat_keparahan Eng. ver: Severity Level
Perilaku	Siswa beberapa kali datang terlambat dan melanggar tata tertib sekolah.	Selesai	Siswa sudah diberikan pembinaan dan menunjukkan perubahan positif.	Tinggi
Eng. ver: Behavioral	Eng. ver: The student has repeatedly arrived late and violated school regulations.	Eng. ver: Completed	Eng. ver: The student has received guidance and is showing positive changes.	Eng. ver: High
Sosial	Terlibat konflik dengan teman sebaya yang menyebabkan ketegangan di kelas.	Selesai	Dilakukan mediasi dan kondisi sosial siswa mulai membaik.	Tinggi
Eng. ver: Social	Eng. ver: Involved in conflicts with peers that caused tension in the classroom.	Eng. ver: Completed	Eng. ver: Mediation was conducted and the student's social condition has begun to improve.	Eng. ver: High
Akademik	Nilai beberapa mata pelajaran menurun dan sering tidak mengumpulkan tugas.	Diproses	Sedang dilakukan pendampingan belajar oleh guru BK.	Sedang
Eng. ver: Academic	Eng. ver: Grades in several subjects have declined and the student frequently fails to submit assignments.	Eng. ver: In Progress	Eng. ver: The student is currently receiving learning support from the counselling teacher.	Eng. ver: Moderate
Sosial-Emosional	Siswa melaporkan merasa tidak nyaman karena tindakan ejekan dari beberapa teman.	Diproses	Kasus sedang dipantau untuk mencegah potensi bullying berlanjut.	Tinggi
Eng. ver: Social-Emotional	Eng. ver: The student reported feeling uncomfortable due to teasing behavior from several peers.	Eng. ver: In Progress	Eng. ver: The case is being monitored to prevent potential bullying from escalating.	Eng. ver: High

Table 3 presents the class distribution of the validated student report dataset. The dataset exhibits a moderate class imbalance, with bullying cases representing 38.8% of the total samples. This distribution reflects real-world school counselling conditions, where not all student reports involve bullying incidents (Y. Zhang et al., 2023). Table 3 summarizes the final validated dataset, where all 1,000 reports retained after the validation process were used for class distribution analysis.

Table 3. Class distribution of validated student report dataset

Label	Description	Number of Samples	Percentage
0	Non-bullying	612	61.2%
1	Bullying	388	38.8%
Total		1,000	100%

2.1.3. Pseudo-Labelling and Validation Process

The pseudo-labelling process was conducted in three stages. First, the pretrained Hybrid CNN-LSTM model generated initial labels for all unlabeled student reports. Second, two school counsellors (guidance and counselling teachers) with professional experience in handling student bullying cases independently reviewed all 1,000 reports and assigned binary labels (bullying or non-bullying) without access to each other's annotations. Prior to data filtering, inter-annotator agreement was evaluated using Cohen's Kappa, yielding a κ value of 0.82, which indicates strong agreement beyond chance between the two annotators (Rau & Shih, 2021). All 1,000 reports reached full agreement between the annotators; therefore, no samples were discarded at the inter-annotator validation stage.

In the third stage, the agreed human annotations were compared with the model-generated pseudo-labels. Only samples for which both annotators' labels were consistent with the model prediction were retained for training, while inconsistent or ambiguous cases were excluded. As a result, 742 reports were accepted, and 258

reports were rejected. Only validated samples were included in the final fine-tuning dataset to ensure label reliability in this low-resource setting.

2.2. Data Gathering

Student reports are collected through a web-based anonymous reporting portal. All personally identifiable information, including names, locations, and specific dates, is automatically removed using NLP-based entity detection and content masking techniques. This process follows established ethical guidelines for digital research involving minors and sensitive user-generated content. In addition to the student reports dataset, publicly available textual data are collected from open-source repositories and social media platforms for the pretraining phase.

These public datasets contain no personally identifiable information and are accessed in compliance with platform usage policies and open-data licenses. The resulting student reports dataset represents a low-resource NLP setting, reflecting real-world constraints in educational institutions. To support contextual understanding, aggregated and anonymized counselling-unit logs are also utilized to analyze reporting frequency and submission patterns without compromising individual privacy.

2.3. Model Architecture

The proposed model employs a Hybrid CNN-LSTM architecture, as illustrated in Figure 2, where convolutional layers first extract local lexical and n-gram features related to bullying expressions (e.g., slang, insults, and abusive terms), followed by LSTM layers that model long-range dependencies and narrative structure across sentences. This sequential design enables the model to capture both surface-level linguistic cues and deeper contextual information, which is essential for interpreting extended and emotionally expressive student reports.

The detailed architectural configurations for both the pretraining and fine-tuning stages are summarized in Table 4. While the overall model structure remains consistent across stages, key parameters such as embedding dimensionality, number of CNN filters, and LSTM units are adjusted during fine-tuning to reduce model complexity and mitigate overfitting under low-resource conditions.

Unlike a conceptual workflow diagram, Figure 2 illustrates the layer-level technical architecture of the proposed Hybrid CNN–LSTM model. The diagram explicitly represents the sequential flow from tokenized input, embedding representation, convolutional feature extraction, pooling, and LSTM-based contextual modeling to the final sigmoid classification layer. Detailed hyperparameter configurations for each component are provided in Table 4 to ensure reproducibility. (Raj et al., 2021; Ullah et al., 2024).

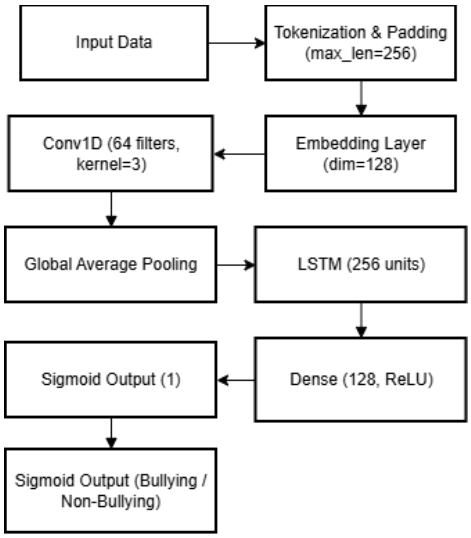

Figure 2. Architecture of the proposed Hybrid CNN-LSTM model

As shown in Figure 2 and detailed in Table 4, the same Hybrid CNN-LSTM architecture is applied during both training stages, ensuring architectural consistency while allowing computationally efficient domain adaptation. This design choice is supported by prior studies on Indonesian-language cyberbullying detection, which report that CNN-LSTM models consistently outperform standalone CNN or LSTM architectures, particularly for long and noisy textual data (Raj et al., 2021; Ullah et al., 2024; Xu, Song, et al., 2024).

Table 4. Model components for pretraining and fine-tuning

Component	*Pretraining* (*Public data)*	*Fine-tuning* (Real Data)	Description
Sequence length	256	256	Adapting to the student's language style
Embedding dim.	300	128	Simplified
Char embedding	64	32	Efficiency
LSTM layer	LSTM (256×2)	LSTM (256×1)	Local features
CNN filters (Conv1D)	128	64	Local n-gram feature extraction
Pooling	Global Max	Global Avg	Adjusting the distribution
Dense	256	128	Regulation
Total params	11.5M	4.2M	Fine-tuned version is comparatively lighter

The reduced configuration during fine-tuning reflects an intentional trade-off between model capacity and generalization ability under low-resource conditions. This design minimizes overfitting while preserving the core representations learned during large-scale pretraining.

Although Transformer-based models such as IndoBERT and large language models (LLMs) have demonstrated strong performance in various NLP tasks, they were not adopted in this study for three reasons. First, Transformer models typically require substantially larger computational resources, which limits deployment feasibility in school counselling systems. Second, prior studies have shown that CNN-LSTM architectures remain competitive for long, noisy, and emotionally expressive texts, particularly in low-resource Indonesian settings (Raj et al., 2021; Xu, Jing, et al., 2024). Third, LLM-based approaches raise additional concerns related to data privacy, interpretability, and ethical governance in educational environments involving minors (Crompton & Burke, 2023) Therefore, a Hybrid CNN-LSTM model offers a balanced trade-off between performance, interpretability, and practical deployment constraints.

2.4. Dataset Partitioning

The complete dataset consists of 82,694 samples and is partitioned into training, validation, and test sets using a stratified sampling strategy. An 80:10:10 split is applied to preserve class distribution across all subsets, as shown in Table 5. Stratified partitioning ensures proportional representation of both majority (non-bullying) and minority (bullying) classes, reducing evaluation bias and supporting reliable generalization assessment (Yu & Zhao, 2021; S. Zhang et al., 2024; Y. Zhang et al., 2023).

Table 5. Dataset distribution across training, validation, and test sets

Subset	Total Sample	Proportion of label non-bullying (0)	Proportion of label Bully (1)	Total Proportion
Training	66.155	71.5%	28.5%	80%
Validation	8.269	71.5%	28.5%	10%
Test	8.270	71.5%	28.5%	10%
Total	82.694	-	-	100%

2.5. Training Strategy

A two-stage training strategy is employed, consisting of a pretraining phase and a fine-tuning phase. The model is first pretrained for 10 epochs using large-scale publicly available textual data to learn general linguistic patterns and abusive-language representations (Chen et al., 2021; Rahamim et al., 2022; Yan et al., 2023).

Subsequently, the pretrained model is fine-tuned for 5 epochs on real student counselling reports to adapt the learned representations to school-specific language characteristics and contextual nuances (Y. Zhang et al., 2023).

The hyperparameter configurations for both training stages are summarized in Table 6. During fine-tuning, a lower learning rate is applied to enable gradual domain adaptation while preserving pretrained knowledge, following standard transfer-learning practices (Afriyani et al., 2024; Crompton & Burke, 2023; Yan et al., 2023). Early stopping with a patience of three epochs is applied exclusively during the fine-tuning stage to mitigate overfitting on the limited labeled dataset. All experiments are conducted using a fixed random seed (42) across TensorFlow, NumPy, and Python environments to ensure reproducibility (Christian et al., 2024). Evaluation metrics, including accuracy, precision, recall, and F1-score, are computed using standard implementations from the Scikit-learn library (Handayani et al., 2025; X. Yang et al., 2023).

Table 6. Model configuration and hyperparameter settings

Parameter	*Pretraining* (*Public data*)	*Fine-tuning* (Real Data)	Description
Optimizer	Adam	Adam	Stable for deep learning
Learning Rate	0.001	0.0001	Decreased during adaptation
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy	Binary classification
Batch Size	32	32	Consistent
Epoch	10	5	Early stopping applied
Dropout	0.5	0.5	Regularisation
L2 Regularization	0.001	0.001	Weight stabilisation
Early Stopping	-	Patience = 3	Prevents overfitting

2.6. Model Evaluation

Model performance is evaluated using accuracy, precision, recall, and F1-score. These metrics capture both overall classification performance and sensitivity to bullying cases, where false negatives can have serious practical implications (Christian et al., 2024). A confusion matrix analysis is further conducted to examine error patterns, including false positives and false negatives. In addition, the proposed model is benchmarked against classical baseline classifiers, such as Logistic Regression and Support Vector Machines, following standard NLP evaluation practices (Handayani et al., 2025; L. Yang et al., 2023).

3. Result and Discussion

The pretraining stage was conducted using publicly available data designed to represent general characteristics of bullying-related text. During this stage, the Hybrid CNN-LSTM model indicated its initial capability to identify linguistic patterns associated with aggressive behavior. The model achieved a pre-fine-tuning accuracy of 0.9225, while the training loss gradually decreased from 0.491 to 0.462, indicating stable learning behavior and effective pattern generalization.

These findings align with the domain-adaptive pretraining framework proposed by (Chen et al., 2021; Yan et al., 2023), which emphasizes the importance of pretraining on domain-relevant data to build robust linguistic representations prior to downstream adaptation. In the context of this study, the pretraining results suggest that the model learned recurrent aggression-related structures, including insults, intimidation, and threatening expressions, thereby establishing a strong foundation for subsequent fine-tuning.

3.1. Fine-Tuning Results

Following pretraining, the model was fine-tuned using real-world bullying reports authored by Indonesian students. This stage is critical due to the substantial linguistic variability present in student-generated text, including non-standard grammar, abbreviations, code-mixing, emotionally expressive language, and culturally contextualized expressions. As summarized in Table 7, the proposed model achieved an accuracy of 0.8698, a precision of 0.7415, a recall of 0.8570, and an F1-score of 0.7951. The results indicate stable classification. The observed performance is consistent with findings reported by (Akella et al., 2022) which highlight the effectiveness of two-stage training strategies combining large-scale pretraining with domain-specific fine-tuning.

To provide a more robust evaluation, 95% confidence intervals (CI) were computed for accuracy, precision, recall, and F1-score using bootstrapping with 1,000 resamples. As shown in Table 7, the relatively narrow confidence intervals indicate stable performance and limited variance across evaluation runs, suggesting reliable generalization of the proposed model.

Table 7. Evaluation results after fine-tuning

Metric	Value	95% CI
Accuracy	0.8698	[0.851, 0.888]
Precision	0.7415	[0.712, 0.771]
Recall	0.8570	[0.834, 0.879]
F1-score	0.7951	[0.768, 0.822]

The results further indicate that the Hybrid CNN-LSTM model outperforms conventional machine learning approaches by effectively integrating local lexical pattern extraction (via CNN) with contextual and emotional sequence modeling (via LSTM). This suggests that narrative educational texts require deeper contextual modeling beyond surface-level lexical matching.

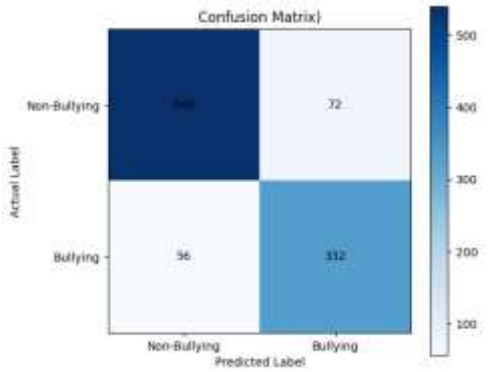

Figure 3. Confusion matrix on the validated real student counselling report dataset (n = 1,000).

Figure 3 illustrates the confusion matrix of the proposed Hybrid CNN-LSTM model evaluated on the validated student counselling report dataset. Out of 388 bullying reports, the model correctly identified 332 instances, achieving a recall of 85.7%, while 56 cases were misclassified as non-bullying. For the non-bullying class, 540 out of 612 reports were correctly classified, with 72 instances falsely predicted as bullying. These results indicate that the model prioritises sensitivity in detecting bullying-related content, resulting in a relatively low number of false negatives, which is particularly important for early-warning and intervention systems in school counselling contexts (Barrios-Cogollo et al., 2025). Overall, the results indicate that the integration of pretraining and fine-tuning enables the Hybrid CNN-LSTM model to detect both explicit and implicit bullying expressions embedded in natural student language, while providing interpretable performance through absolute numbers and row-wise percentages.

Although the proposed model achieves high recall, its precision is comparatively lower. This indicates a tendency to generate false positives, reflecting a recall-oriented classification strategy. In the context of school-based bullying detection, this trade-off is considered acceptable, as missing actual bullying cases may lead to more severe ethical and psychological consequences than generating additional alerts that can be reviewed by school counselors. Nevertheless, this behavior highlights the need for future improvements in false-positive reduction, such as threshold optimization or human-in-the-loop validation mechanisms.

3.2. Comparison with Classic Method

Table 8 learns a comparative evaluation of the proposed Hybrid CNN-LSTM model against Logistic Regression and Linear SVM across different report-length categories. For short texts, both classical models achieve relatively strong performance, with identical F1-scores of 0.80. This result indicates that explicit bullying expressions commonly found in short reports can be effectively captured using surface-level lexical features and linear decision boundaries, which are sufficient when contextual complexity is limited (Handayani et al., 2025; L. Yang et al., 2023).

Table 8. Performance comparison across different text length categories

Model	Short F1 (≤50 words)	Medium F1 (51-150 words)	Long F1 (>150 words)
Logistic Regression	0.80	0.62	0.58
SVM Linear	0.80	0.63	0.58
HYBRID CNN-LSTM	0.86	0.85	0.81

However, the performance of both classical models declines substantially as report length increases, with noticeable reductions in F1-score for medium and long texts. This degradation suggests that linear classifiers struggle to capture long-range contextual dependencies, implicit emotional expressions, and narrative structures inherent in extended student reports. In contrast, the proposed Hybrid CNN-LSTM maintains stable and consistently higher performance across all text-length categories, achieving particularly strong results on medium and long reports. Figure 4 provides a visual comparison of F1-scores across report-length categories, illustrating the performance gap between classical models and the Hybrid CNN-LSTM as text length increases.

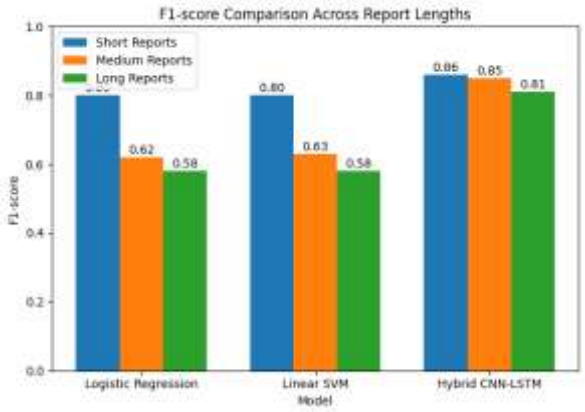

Figure 4. F1-score comparison across report-length categories

For baseline comparisons, Logistic Regression and Linear SVM were implemented using TF-IDF feature representations with unigram and bigram configurations. Term frequency-inverse document frequency weighting was applied, and the feature space was restricted to the top 10,000 most frequent unigram and bigram terms to control dimensionality and sparsity. The performance degradation observed in Logistic Regression and Linear SVM on long narrative texts reflects the limitations of static lexical representations and linear decision boundaries in capturing long-range contextual dependencies, event chronology, and implicit emotional cues. (Yu & Zhao, 2021; S. Zhang et al., 2024; Y. Zhang et al., 2023) reported that linear classifiers relying on surface-level lexical features tend to underperform on educational texts containing emotional progression and culturally contextualized expressions, particularly when bullying indicators are implicitly embedded within narratives. In contrast, the sequential architecture of the Hybrid CNN-LSTM enables effective modeling of temporal and contextual relationships across sentences, making it more suitable for narrative-based bullying detection tasks. This observation is further supported by (Maragheh et al., 2024), who indicated that sequential deep-learning architectures consistently outperform linear baselines in modeling extended textual sequences due to their ability to capture inter-sentence dependencies and emotional continuity.

Beyond descriptive performance metrics, statistical significance testing was conducted to further validate the observed performance differences. A McNemar test was applied, as it is appropriate for paired nominal data where multiple models are evaluated on the same test instances. The results indicate that the proposed Hybrid CNN-LSTM model significantly outperforms both Logistic Regression and Linear SVM ($p < 0.05$), confirming that the observed performance gains are statistically significant and not attributable to random variation.

3.3. User Study

The usability study involved 2 school counselors (Guru BK), 3 school teachers, and 20 students, resulting in a total of 25 participants. This sample size meets the minimum requirements for usability evaluation using the System Usability Scale (SUS), which has been shown to produce reliable results even with moderate participant numbers (Brooke, 2020). The SUS questionnaire consists of 10 standardized statements rated on a five-point Likert scale ranging from “strongly disagree” to “strongly agree.” The reliability of the instrument was assessed using Cronbach’s alpha, yielding a value of $\alpha = 0.60$, which indicates an acceptable level of internal consistency for usability assessment instruments such as SUS.

Table 9. Stakeholder evaluation results

Group	Mean Score	Category
Educators	72.5	Good
Students	73.25	Good
Total	72.875	Good

3.4. Main Discussion

The results indicate that the performance of the proposed bullying detection system is driven by the interaction between the Hybrid CNN-LSTM architecture, the use of public data for pretraining, and the linguistic

characteristics of Indonesian student reports. Together, these components enable the model to detect both explicit and implicit bullying expressions embedded in narrative counselling reports. The effectiveness of the Hybrid CNN-LSTM architecture can be attributed to its ability to capture both local lexical cues and long-range contextual dependencies. Prior studies have shown that combining convolutional and recurrent layers enhances a model's capacity to represent narrative text with complex temporal structure (Y. Zhang et al., 2023). This capability aligns closely with the characteristics of student bullying reports, which are often long, emotional, and chronologically structured.

Although the student reports dataset originates from a limited number of schools, this design choice is methodologically justified given the ethical and practical constraints associated with collecting sensitive educational data. Rather than emphasizing data volume, the proposed framework prioritizes robustness under low-resource conditions. Pretraining on public data allows the model to acquire foundational linguistic representations related to bullying before domain-specific adaptation. This strategy has been shown to improve model stability and generalization in low-resource settings. Fine-tuning on real student reports further enables the model to adapt to the linguistic, cultural, and emotional nuances characteristic of Indonesian educational contexts, supporting findings on the importance of domain-specific adaptation in educational NLP applications (Yan et al., 2023).

4. Conclusion

This study presents a Hybrid CNN-LSTM model for detecting bullying in narrative student counselling reports, effectively capturing both local lexical patterns and contextual-emotional sequences. While the model shows strong performance, limitations include a dataset from few schools, focus on a single language (Indonesian), and relatively lower precision. Future work should expand and diversify the dataset, explore multilingual applications, integrate explainable AI (XAI), and evaluate performance in varied school contexts to enhance generalizability and interpretability. This study did not conduct an ablation analysis due to the limited size of the labeled student counselling dataset, as such experiments could lead to unstable conclusions and are therefore left for future work.

References

Afriyani, S., Surono, S., & Solihin, I. M. (2024). Chi-Square Feature Selection with Pseudo-Labelling in Natural Language Processing. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *8*(3), 896. https://doi.org/10.31764/jtam.v8i3.22751

Akar, F. (2024). Performance Analysis of NLP-Based Machine Learning Algorithms in Cyberbullying Detection. *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *17*(2), 445–459. https://doi.org/10.18185/erzifbed.1474112

Akella, P., Ahmadi, M., & Ames, A. D. (2022). *A Scenario Approach to Risk-Aware Safety-Critical System Verification*. http://arxiv.org/abs/2203.02595

Alhakim, A., Meriana, A., Besley, B., & Khoesasi, W. (2022). Pengaruh Bullying Dan Hate Speech Terhadap Kesehatan Mental Remaja Di SMK Yehonala. *Prosiding National Conference for Community Service Project (NaCosPro)*, *4*(1), 104–114. https://doi.org/10.37253/nacospro.v4i1.6925

Andini, M. (2025). *Indonesian Multi-Source Bullying & Cyberbullying Datase*. Github. https://github.com/muthiaandinini/Bullying/blob/main/Readme.md

Barrios-Cogollo, C., Gómez Gómez, J., & De-La-Hoz-Franco, E. (2025). Comparative Analysis of Classification Models for Cyberbullying Detection in University Environments. *Applied Sciences (Switzerland)*, *15*(18). https://doi.org/10.3390/app151810100

Brooke, J. (2020). SUS: A Quick and Dirty Usability Scale. In P. W. ; T. B. ; M. I. L. ; W. B. A. Jordan (Ed.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis. https://doi.org/10.1201/9781498710411-35

Chen, Q., Zhu, Y., & Chui, W. H. (2021). A Meta-Analysis on Effects of Parenting Programs on Bullying Prevention. In *Trauma, Violence, and Abuse* (Vol. 22, Number 5, pp. 1209–1220). SAGE Publications Ltd. https://doi.org/10.1177/1524838020915619

Christian, Y., Wibowo, T., & Lyawati, M. (2024). Sentiment Analysis by Using Naïve Bayes Classification and Support Vector Machine, Study Case Sea Bank. *Sinkron*, *9*(1), 258–275. https://doi.org/10.33395/sinkron.v9i1.13141

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00392-8

Hafiza, A. A., & Setiawan, E. B. (2025). Enhancing Cyberbullying Detection on Platform "X" Using IndoBERT and Hybrid CNN-LSTM Model. *Jurnal Teknik Informatika (Jutif)*, *6*(2), 655–672. https://doi.org/10.52436/1.jutif.2025.6.2.4321

Hamapu, A. (2024, March 3). Polisi Ungkap Motif Pelaku Bully Remaja di Batam: Sakit Hati-Saling Ejek. *DetikNews*. https://news.detik.com/berita/d-7222364/polisi-ungkap-motif-pelaku-bully-remaja-di-batam-sakit-hati-saling-ejek

Handayani, S., Isnanto, R., & Warsito, B. (2025). Co-training pseudo-labeling for text classification with support vector machine and long short-term memory. *IAES International Journal of Artificial Intelligence*, *14*(3), 2158–2168. https://doi.org/10.11591/ijai.v14.i3.pp2158-2168

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). *A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios*. https://aclanthology.org/2021.naacl-main.201/

KPAI. (2023, November 29). Rakornas dan Ekspose KPAI 2023, Membangun Indonesia Bebas Kekerasan Terhadap Anak. *KPAI*. https://www.kpai.go.id/publikasi/rakornas-dan-ekspose-kpai-2023-membangun-indonesia-bebas-kekerasan-terhadap-anak

Le, H. T. H., Tran, N., Campbell, M. A., Gatton, M. L., Nguyen, H. T., & Dunne, M. P. (2019). Mental health problems both precede and follow bullying among adolescents and the effects differ by gender: A cross-lagged panel analysis of school-based longitudinal data in Vietnam. *International Journal of Mental Health Systems*, *13*(1). https://doi.org/10.1186/s13033-019-0291-x

Maragheh, H. K., Gharehchopogh, F. S., Majidzadeh, K., & Sangar, A. B. (2024). A Hybrid Model Based on Convolutional Neural Network and Long Short-Term Memory for Multi-label Text Classification. *Neural Processing Letters*, *56*(2). https://doi.org/10.1007/s11063-024-11500-8

Purba, M., Paisal, P., Pambudi Darmo, C., Noprisson, H., & Ayumi, V. (2024). Model Of Indonesian Cyberbullying Text Detection Using Modified Long Short-Term Memory. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *10*(1), 9–14. https://doi.org/10.33480/jitk.v10i1.5239

Rahamim, A., Uziel, G., Goldbraich, E., & Anaby-Tavor, A. (2022). Text Augmentation Using Dataset Reconstruction for Low-Resource Classification. *Findings of the Association for Computational Linguistics: ACL 2023*, 7389–7402.

Raj, C., Agarwal, A., Bharathy, G., Narayan, B., & Prasad, M. (2021). Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques. *Electronics (Switzerland)*, *10*(22). https://doi.org/10.3390/electronics10222810

Rau, G., & Shih, Y.-S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes*, *53*, 101026. https://doi.org/https://doi.org/10.1016/j.jeap.2021.101026

Setiawan, Y., Ulva Maulidevi, N., Surendro, K., & Korespondensi, P. (2022). Deteksi Cyberbullying Dengan Mesin Pembelajaran Klasifikasi (Supervised Learning): Peluang Dan Tantangan. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *9*. https://doi.org/10.25126/jtiik.202296747

Ullah, K., Ahsan, M., Hasanat, S. M., Haris, M., Yousaf, H., Raza, S. F., Tandon, R., Abid, S., & Ullah, Z. (2024). Short-Term Load Forecasting: A Comprehensive Review and Simulation Study with CNN-LSTM Hybrids Approach. *IEEE Access*, *12*, 111858–111881. https://doi.org/10.1109/ACCESS.2024.3440631

UNICEF. (2021, June 25). Indonesia: Hundreds of children and young people call for kindness and an end to bullying. *UNICEF*. https://www.unicef.org/indonesia/press-releases/indonesia-hundreds-children-and-young-people-call-kindness-and-end-bullying?

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., Casadio, R., Capella-Gutierrez, S., Cirillo, D., Del Conte, A., Dimopoulos, A. C., Del Angel, V. D., Dopazo, J., Fariselli, P., Fernández, J. M., Huber, F., Kreshuk, A., Lenaerts, T., Martelli, P. L., … Tosatto, S. C. E. (2021). DOME: recommendations for supervised machine learning validation in biology. *Nature Methods*, *18(10)*, 1122–1127. https://doi.org/10.1038/s41592-021-01205-4

Xu, P., Jing, L., & Yu, J. (2024). *Enhancing Multi-Label Text Classification under Label-Dependent Noise: A Label-Specific Denoising Framework*. https://doi.org/10.18653/v1/2024.findings-emnlp.324

Xu, P., Song, M., Liu, L., Liu, B., Sun, H., Jing, L., & Yu, J. (2024). Noisy Multi-Label Text Classification via Instance-Label Pair Correction. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1446–1458). Association for Computational Linguistic. https://doi.org/10.18653/v1/2024.findings-naacl.93

Yan, W., Yuan, Y., Yang, M., Zhang, P., & Peng, K. (2023). Detecting the risk of bullying victimization among adolescents: A large-scale machine learning approach. *Computers in Human Behavior*, *147*. https://doi.org/10.1016/j.chb.2023.107817

Yang, L., Huang, B., Guo, S., Lin, Y., & Zhao, T. (2023). A Small-Sample Text Classification Model Based on Pseudo-Label Fusion Clustering Algorithm. *Applied Sciences (Switzerland)*, *13*(8). https://doi.org/10.3390/app13084716

Yang, X., Song, Z., King, I., & Xu, Z. (2023). A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8934–8954. https://doi.org/10.1109/TKDE.2022.3220219

Yu, S., & Zhao, X. (2021). The negative impact of bullying victimization on academic literacy and social integration: Evidence from 51 countries in PISA. *Social Sciences and Humanities Open*, *4*(1). https://doi.org/10.1016/j.ssaho.2021.100151

Yuliandra, R. (2025). Sepanjang 2025, Kasus Kekerasan terhadap Perempuan dan Anak di Batam Capai 141 Kasus. In *Batam Pos*. https://batampos.co.id/2025/06/02/sepanjang-2025-kasus-kekerasan-terhadap-perempuan-dan-anak-di-batam-capai-141-kasus/

Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-024-00467-0

Zhang, Y., Jiang, M., Meng, Y., Zhang, Y., & Han, J. (2023). PIEClass: Weakly-Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12655. https://doi.org/10.18653/v1/2023.emnlp-main.780