Journal of Theoretical and Applied Information Technology
15st June 2026. Vol.104. No.11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CYBERBULLYING SEVERITY DETECTION IN MALAY-ENGLISH CODE-MIXED TEXT: A HYBRID NLP-CNN APPROACH WITH USER BEHAVIOR PROFILING

IBRAHIM INUSA[1], RINA MD ANWAR[2], ASMIDAR ABU BAKAR[3], FIZA ABDUL RAHIM[4], MARINA MD DIN[5], ALIZA ABDUL LATIF[6]

[1] Postgraduate Research Assistant, Universiti Tenaga Nasional, College of Computing and Informatics, 43009 Kajang, Malaysia
[2, 5, 6] Senior Lecturer, Universiti Tenaga Nasional, College of Computing and Informatics, 43009 Kajang, Malaysia
[3] Associate Professor, Universiti Tenaga Nasional, College of Computing and Informatics, 43009 Kajang, Malaysia
[4] Senior Lecturer, Universiti Teknologi Malaysia, Department of Computing, 54100 Kuala Lumpur, Malaysia
E-mail: [1]talk2ajmilala@gmail.com, [2]mrina@uniten.edu.my, [3]asmidar@uniten.edu.my, [4]fiza.abdulrahim@utm.my, [5]marina@uniten.edu.my, [6]aliza@uniten.edu.my

ABSTRACT

Cyberbullying detection remains a significant and ongoing challenge, particularly within multilingual code-mixed discourse such as Malay–English, where existing approaches remain limited to binary identification and are predominantly developed on monolingual datasets, overlooking both the varying severity of cyberbullying and user behavioral patterns, which are essential for developing effective interventions. Despite the growing prevalence of cyberbullying, severity-aware detection for Malay–English code-mixed text remains largely underexplored, leaving a methodological gap between real-world multilingual communication patterns and current automated moderation systems. Given the linguistic variability of code-mixed discourse and the need to capture both lexical intensity and contextual semantic cues in Malay–English code-mixed cyberbullying interactions, a hybrid modeling strategy that integrates statistical lexical features and neural contextual representations provides a more suitable framework for severity-aware cyberbullying detection. To address this limitation, this study proposes a hybrid Natural Language Processing–Convolutional Neural Network (NLP–CNN) framework designed to classify cyberbullying instances across three severity levels (low, medium, and high) while simultaneously profiling user behavioral tendencies. A balanced Malay–English cyberbullying corpus comprising 52,140 annotated instances was constructed through multi-source dataset integration and standardized labeling to support model training and evaluation. The proposed architecture integrates TF-IDF (Term Frequency–Inverse Document Frequency) lexical representations with CNN-based contextual feature learning, enabling the model to capture both surface-level linguistic cues and deeper semantic patterns characteristic of code-mixed discourse. Experimental evaluation demonstrates strong predictive performance, achieving 98.42% classification accuracy with consistently high precision, recall, and F1-scores across severity categories. Ablation analysis further shows that neither lexical nor neural representations alone sufficiently capture cyberbullying severity, whereas their hybrid integration yields more balanced classification outcomes. Beyond predictive performance, a rule-based behavioral profiling module enhances interpretability by mapping severity predictions to distinct interaction archetypes. These findings demonstrate that combining hybrid deep-learning architectures with interpretable behavioral analysis provides a scalable and context-sensitive approach for cyberbullying severity detection in multilingual code-mixed environments, supporting more effective automated moderation and digital safety interventions.

Keywords: *Convolutional Neural Networks, Cyberbullying Detection, Malay–English Code-Mixing, Natural Language Processing, Severity Classification, User Behavior Profiling*

1. INTRODUCTION

Cyberbullying is defined as a deliberate and repeated use of digital technology to intimidate, harass, or harm others [2]. It also encompasses a range of behaviors, including verbal attacks, spreading rumors, and sharing malicious content [1]. The growing prevalence of cyberbullying has emerged as a global concern, largely because of its profound psychological, social, and legal repercussions [3]. This further becomes even more complex within multilingual code-switching online spaces, where frequent code-switching and mixed-language expressions complicate the cyberbullying detection and development of effective interventions [3], [7].

The widespread adoption of online social networks (OSNs) has made online communication an integral part of daily life, while simultaneously intensifying the prevalence of cyberbullying [8]. This growing prevalence extends to multilingual code-switch discourse such as Malay-English that characterizes online discourse in Southeast Asia, according to the Microsoft Global Youth Online Behavior Survey, Malaysia was ranked 17th out of 25 countries on cyberbullying practices. A similar survey by IPSOS (Institut Public de Sondage d'Opinion Secteur) Global Advisor ranked sixth out of 28 countries globally and second in Asia [38], [55]. Between January and June 2020 alone, the Malaysian Communications and Multimedia Commission (MCMC) logged 11,235 complaints involving digital offenses such as cyberbullying [8]. This followed a period between 2016 and late 2020 where 10,406 reports were filed, and a separate MCMC assessment from 2018 to 2020 flagged 3,762 specific cyberbullying instances [56].

This digital aggression poses serious risks to both mental and physical well-being, having been linked to anxiety, depression, emotional distress, and suicide [55]. The escalation from online harassment to physical harm is evidenced by high-profile cases: the 2020 suicide of 20-year-old Thivya Nayagi Rajendran following TikTok harassment; the 2023 suicide attempt by actress Lee Yuan Ling due to online violence; and the 2023 fatal stabbing of singer Yuki Koh by an obsessive fan who had stalked her on social media [57]. Beyond individual mental health degradation, online abuse has broader societal implications, with the potential to incite real-world violence and riots [8].

The Malay-English code-switching, often referred to as "Manglish" or "Bahasa Rojak" (mixed language), is characterized by relaxed grammar rules and colloquial forms that disregard standard orthography [58]. This phenomenon, similar to Spanglish, has speakers primarily concentrated in Malaysia, Indonesia, Brunei, and Singapore [38]. It involves complex "intracentric code conversion" and "congruent lexicalization," where words from both languages are inserted into a single frame, blurring language boundaries [59], [60]. Cyberbullying detection in Malay-English code-switching primarily employs binary classification (identifying "bullying" vs. "Non-bullying") or sentiment-based approaches (Positive/Neutral/Negative). While effective for coarse screening, these approaches are inherently limited in their ability to capture variations in harm intensity, thereby constraining their usefulness for real-world risk assessment [56], [61]. In response, severity-based detection has emerged as a more informative paradigm, enabling differentiation across multiple levels of cyberbullying harm rather than treating abuse as a uniform phenomenon [11], [19].

While significant efforts have been made to detect cyberbullying through text analysis, there remains a significant gap in the assessment of the severity of cyberbullying instances [7], [8], especially in multilingual code-mixed text, such as Malay-English [9], [10]. However, detecting the severity of cyberbullying is crucial in designing intervention strategies because it reveals how serious or harmful a specific cyberbullying incident may be, facilitating appropriate mitigation measures at all levels of cyberbullying behavior because not all types of cyberbullying are the same, as some types are less harmful (low severity) while others are extremely dangerous and could lead to serious mental health issues or even suicide (high severity) [11], while some fall in between these extremes [12]. Thus, attention and interventions need to be made according to the degree of severity in order to lessen the influence of cyberbullying [13], [14].

Against this background, it becomes necessary to move beyond conventional binary cyberbullying detection toward approaches that capture the varying intensity and behavioral nature of harmful online interactions, particularly within multilingual code-mixed environments such as Malay–English discourse. Existing models, largely developed for monolingual datasets, remain limited in their ability to represent the linguistic variability and contextual ambiguity present in such settings. Consequently, this study is motivated by the need to

develop a severity-aware computational framework that also accounts for user behavioral tendencies, enabling a more context-sensitive understanding of cyberbullying dynamics and supporting more effective intervention strategies in multilingual online communication spaces.

1.1 Current Applications of Related Technology

Recent advances in natural language processing (NLP) and machine learning have enabled a wide range of automated cyberbullying detection systems deployed across social media platforms, primarily for large-scale content moderation and early risk identification [50]. Early implementations predominantly relied on supervised machine learning models such as Support Vector Machines, Naïve Bayes, and Logistic Regression using lexical representations (e.g., n-grams and TF-IDF), demonstrating effectiveness in identifying explicit abusive language [21], [56]. These systems are widely applied in monitoring offensive content, hate speech, and user-generated interactions, particularly where linguistic signals such as profanity, insults, and aggressive expressions serve as primary indicators of Cyberbullying [16].

Subsequent developments have incorporated deep learning architectures, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) models, to capture contextual semantics and sequential dependencies in online discourse [18]. These models have been applied in large-scale social media analytics, enabling improved detection of implicit aggression, sarcasm, and context-dependent abusive patterns that are often missed by traditional approaches [34]. More recent studies further integrate linguistic, behavioral, and contextual features—such as user interaction patterns, message frequency, and engagement dynamics—to enhance detection robustness and reflect real-world cyberbullying behaviors [14], [37].

Despite these advancements, current applications remain largely constrained to binary classification frameworks and monolingual datasets, with limited operationalization of severity and minimal consideration of multilingual or code-mixed environments. Existing systems, while effective for coarse-grained detection, are not designed to differentiate levels of harm or capture the nuanced linguistic variability present in code-switching contexts such as Malay–English discourse. This limitation restricts their applicability in real-world moderation scenarios where the intensity and contextual meaning of cyberbullying are critical for proportionate intervention.

1.2 Research Problem and Objectives

Cyberbullying severity detection in code-mixed Malay-English text remains insufficiently investigated despite sustained progress in general cyberbullying detection. Existing studies predominantly focus on monolingual settings, thereby limiting their applicability to multilingual code-switching contexts. Prior works, including those by Talpur and O'Sullivan (2020), Vyawahare and Govilkar (2022), Aggarwal et al. (2020), Huang et al. (2023), Prama et al. (2025), and Makkala et al. (2025), propose diverse modeling strategies; however, these approaches are developed and evaluated primarily on single-language corpora and do not explicitly address severity detection.

To address this research gap, our study proposes a hybrid NLP-CNN-based classification framework for cyberbullying severity detection in Malay-English code-switching text. The model classifies cyberbullying instances into three severity levels: low, medium, and high, leveraging NLP techniques for fine-grained linguistic feature extraction and a Convolutional Neural Network (CNN) for hierarchical pattern learning. Beyond severity assessment, the framework further profiles user behavior into four archetypes: impulsive, provocative, harassing, and defensive within each severity level. This dual-level modeling strategy enhances both predictive performance and interpretability, enabling a more comprehensive understanding of cyberbullying intensity and user behavioral dynamics in multilingual online interactions.

1.3 Research Questions

a) How can a hybrid NLP-CNN model be designed and optimized to classify the severity of cyberbullying incidents across low, medium, and high levels in Malay-English code-switching text data?
b) In what ways can the integration of user behavior profiling enhance the interpretability and contextual understanding of severity classification outcomes?
c) How accurately does the proposed hybrid NLP-CNN framework perform when evaluated through empirical testing and system-level deployment?

2. LITERATURE REVIEW

Although machine learning has been extensively applied to the detection of cyberbullying and its severity classification, current research continues to face major challenges in processing multilingual data, managing code-mixed language structures, and achieving consistent severity classification. A study by Obaid, M. H., Guirguis, S. K., & Elkaffas, S. M. [19], [12] utilized LSTM and fuzzy logic to classify cyberbullying severity in English-based tweets, achieving high accuracy (93.67%). However, the study's focus on English data limits its applicability in multilingual settings like Southeast Asia, where Malay-English code-switching is prevalent. With this, prior studies highlight the need for advanced methods and techniques to address linguistic and dataset challenges. Also, Mahmud, S., Ali, A., & Khusro, M. [14] emphasized the scarcity of annotated datasets in low-resource languages, complicating cyberbullying detection in multilingual settings.

2.1 Cyberbullying Severity Classification

The classification of cyberbullying severity can generally be approached in two main ways, as shown by prior studies. The first approach involves rule-based systems, as demonstrated by Vyawahare and Govilkar [18]. In their study, severity is computed through four key linguistic and contextual features: Profanity Ratio (Rp), Capitalization Ratio (Cw), Sentence Length Ratio (Sl), and Semantic Correlation (SCo). These indicators collectively measure the degree of offensive language, the use of aggressive emphasis (such as excessive capitalization), and whether the bullying is directly targeted at an individual (for instance, through the use of pronouns like "you" or explicit usernames). This method enhances automated detection systems by quantifying both linguistic and contextual indicators of online aggression [18], [19]. However, rule-based systems face significant limitations when applied to code-mixed or multilingual content, as linguistic nuances and cultural contexts are difficult to capture using predefined static rules.

The second approach, manual severity labeling, relies on human judgment to evaluate the psychological impact of online messages. In this method, cyberbullying instances are grouped by topic, and severity levels are manually assigned to each group. These annotated labels are then used to train or automate severity classification models [22], [23], [11]. This approach is exemplified by Talpur and O'Sullivan [11], who employed an annotated harassment dataset originally developed by Rezvan, Shetty, and Mather [24]. The dataset was categorized into five main types of harassment: sexual, racial, appearance-related, intelligence-related, and political.

To conduct their severity assessment experiment, the researchers classified the annotated tweets into four severity levels: low, medium, high, and non-cyberbullying. Specifically, sexual and appearance-related tweets were labeled as high severity, political and racial tweets as medium severity, intelligence-related tweets as low severity, while all tweets originally tagged as "non-cyberbullying" were consolidated under the non-cyberbullying category. This process resulted in a single, unified dataset annotated for cyberbullying severity.

2.2 Hybrid NLP-CNN Approach

Natural Language Processing (NLP) provides the computational foundation for analyzing human language by extracting linguistic, syntactic, and semantic features from text. In cyberbullying detection, NLP methods help to identify offensive speech, implicit aggression, and other context-dependent meanings that are not covered by traditional keyword-based systems [25]. Convolutional Neural Network (CNN) models, also known for their visual pattern recognition, have performed well in textual classification techniques as they can learn the hierarchical semantic structure through convolutional filters [26]. Together, NLP preprocessing and CNN feature learning give rise to a hybrid framework for modeling both surface-level linguistic cues and deep contextual dependencies, and this integration can improve the classification of cyberbullying severity by enhancing the ability of models to identify subtle linguistic variations and emotional intensity among multilingual and code-mixed texts [27]. The overview synthesis of the existing cyberbullying severity detection methods is presented in Table 1.

Table I: Comparative Overview of Existing Cyberbullying Severity Detection Studies.

Study/Year	Severity Definition	Classification Levels	Methodology	Key Strengths	Limitations
[11], 2020	Defined by categorizing cyberbullying tweets based on sensitive topics, namely sexuality, racism, physical appearance, intelligence and politics, and assigning severity levels accordingly.	4 levels: Non-cyberbullying, Low, Medium, High	Supervised machine learning framework using PMI-semantic orientation, embedding, sentiment, and lexicon features applied with Naïve Bayes, KNN, Decision Tree, Random Forest, and Support Vector Machine classifiers.	Integrates multiple linguistic and semantic features; demonstrates feasibility of automated severity prediction	Severity categorization is based on authors' intuition and motivation from literature and is open for other researchers to shuffle categories and assign severity levels differently.
[18], 2022	Quantified by profanity ratio, caps use, sentence length, semantic context	5 levels: Non, Light, Moderate, Severe, Nasty	Rule-based annotation with thresholds	Scalable, interpretable	Not effective in code-mixed (Malay-English) data; lacks linguistic nuance
[54], 2025	Bullying severity is determined using a bullying intensity score computed from emotional polarity of comments, user vulnerability factors, and semantic similarity, followed by min–max normalization to categorize severity.	3 levels: Not Bullying (0–0.33), Mild Bullying (0.34–0.66), Severe Bullying (0.67–1)	LSTM-based deep learning model integrating textual features (emotion, topic, Word2Vec) and user-specific attributes; severity labels generated using a bullying intensity measurement technique.	User-specific and explainable severity modeling; suitable for adaptive intervention systems	Requires rich user metadata; generalizability across languages and code-switching contexts is unverified

As summarized in table 1, existing studies on cyberbullying severity detection remain limited in their exploration of severity-aware modelling. Most approaches rely on monolingual datasets, particularly English-language social media corpora, where severity is primarily inferred from lexical intensity cues or predefined topic categories. While these approaches provide useful benchmarks, they do not adequately capture the linguistic complexity of multilingual online environments where code-switching frequently occurs. In particular, the applicability of these models to Malay–English code-mixed discourse remains insufficiently examined, despite the prevalence of such communication patterns in Southeast Asian social media spaces. Furthermore, prior studies rarely integrate severity classification with behavioral interpretation of users, leaving an important gap in understanding how varying levels of cyberbullying intensity relate to underlying interaction patterns. Consequently, there remains a need for computational approaches capable of performing reliable severity-aware detection in Malay–English code-mixed text while also providing interpretable insights into user behavior dynamics in multilingual online interactions.

3. METHODOLOGY

3.1 Conceptual Framework

This study focuses on Malay-English text-based cyberbullying severity detection and proposes a hybrid NLP–CNN framework to classify abusive content into three severity levels (low, medium, and high), alongside profiling user behavior into four archetypes: impulsive, provocative, harassing, and defensive. The proposed framework extends prior monolingual approaches by Talpur and O'Sullivan [11], Unnava and Parasana [22], and Rahman-Laskar et al. [34] to a bilingual sociocultural setting, thereby improving interpretability and contextual fidelity in severity-aware cyberbullying detection, as illustrated in figure 1.

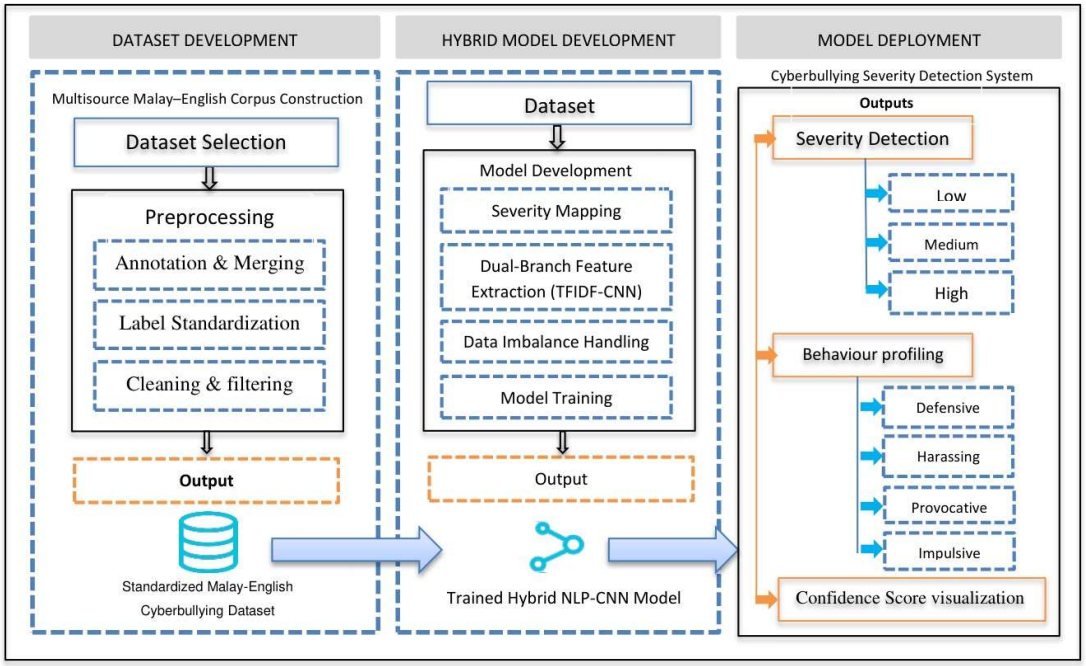

Figure 1: Conceptual framework workflow of the proposed Malay–English cyberbullying severity detection

3.2 Data Collection and Preprocessing

The data collection process initially involved scraping Malay-English interactions from social media platforms such as X (formerly Twitter), using the platform's API to record instances of cyberbullying. However, this method faced limitations due to platform access policies, several ethical considerations, and the significant effort required for manual content annotation. To address these limitations, the research employs a secondary data method using publicly available Malay-English cyberbullying datasets. The selection of datasets was guided by relevance, linguistic composition, and annotation quality to ensure representativeness and reliability. A summary of the compiled datasets is provided in Table 2.

Table 2: Overview of the Malay-English Code-Mixed Cyberbullying Dataset Compilation Process

Ref.	Publication (Title/Year)	Dataset Description	Original Size	Data Included	Actions Performed
[35]	A Bi-Annotated Malay-English Code-switching Dataset of X posts (Formerly Twitter) for Biological Gender Identification and Authorship Attribution, 2024	Code-switched dataset of X posts from 50 Malaysian public figures (25 male, 25 female). Annotated for gender and authorship.	650,409 raw posts	1200	Data manually labeled with AI assistance.
[36]	Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset, 2023	Multi-class dataset of tweets labeled by type (race/ethnicity, gender/sexuality, religion, non-cyberbullying).	99,990 labeled tweets	99,990	Labels standardized for consistency
[37]	A Deep Learning Framework for the Detection of Malay Hate Speech, 2023	Benchmark Malay-language dataset with annotated tweets labeled as hate or non-hate.	4,892 annotated tweets	1888	Non-hate entries removed, Labels standardized

Table 2 provides an overview of the Malay-English code-mixed cyberbullying dataset compilation process adopted in this study. The datasets were compiled from three publicly available sources in Table 2, resulting in an initial pool of 103,077 cyberbullying instances *(1,200 + 99,990 + 1,888)*. These records were consolidated into a unified corpus and stored as *CBDN.csv*. Following the *CBDN.csv* construction, a standardized pre-processing pipeline was implemented, encompassing data cleaning, text normalization, and duplicate removal. This procedure reduced the dataset to 102,171 entries, which were subsequently archived as *CBDN0_preprocessing.csv*. Finally, the non-cyberbullying instances were excluded, resulting in a dataset of 52,140 cyberbullying instances and saved as *CBDN0_FINAL.csv.*

The finalized dataset (*CBDN0_FINAL.csv*), which consists of 52,140 Malay–English cyberbullying instances, was then systematically classified into four categories: *general harassment* (17,990 instances), *gender/sexual harassment* (17,005 instances), *religious bullying* (16,097 instances), and *threat-related abuse* (1,048 instances). This categorical structure provides a coherent basis for severity-level modelling and cross-category analysis. This systematic categorization was grounded in established theoretical frameworks of prior studies, where Threat-related abuse captures explicit or implicit expressions of physical harm or intimidation [13], [11]; gender/sexual harassment comprises degrading or sexualized attacks based on gender or sexuality [42], [28]; religious harassment includes abuse targeting religious identity; and general harassment encompasses non-specific insults or humiliating language [39], [42]. Tables 3 and 4 summarize the manual review, labelling, and standardization procedures, as well as the structured categorization of cyberbullying instances applied during dataset compilation.

Table 3: Dataset Manual Review, Labeling and Standardization Process

Component	Description (Procedure and Purpose)	Ref
Dataset Source	The dataset by Ruhaila Maskat *et al.* (2023) contained 650,409 unlabeled Malay-English social media posts. It's the only dataset to which manual review was applied to.	[38]
Manual Review	Each post was manually inspected to determine if it represented cyberbullying. If identified, it was assigned to one of four predefined categories: *general harassment*, *gender/sexual harassment*, *religious bullying*, or *threat*. Non-offensive posts were labeled as *non-cyberbullying*.	[39][28]
Sample selected from the dataset	A subset of 1,200 manually reviewed posts was selected to construct the final dataset for model training and evaluation.	
AI Assistance (GPT-5)	ChatGPT (GPT-5) was employed to automate the preliminary labeling process using a linguistic prompt based on intent, offensiveness, and target identity.	[40][11]
GPT Output	The model produced initial labels (*Cyberbullying* or *Non-Cyberbullying*) with brief reasoning. This accelerated annotation while maintaining interpretive consistency.	[40]
Human Verification	Human annotator reviewed all GPT-generated labels. Discrepancies were resolved based on linguistic and contextual judgment, ensuring cultural and semantic accuracy.	[41]
Category Standardization (Dataset 2)	Labels from Ahmadinejad & Shahriar (2022) were standardized as follows: *race/ethnicity → general harassment*; *gender/sexuality → gender/sexual harassment*; *religion → religious bullying*; *non-cyberbullying → non-cyberbullying*.	[36]
Category Standardization (Dataset 3)	Labels from Maity *et al.* (2021) were unified by mapping *hate* to *general harassment*, ensuring consistent semantic labeling across sources.	[8]

3.3 Dataset Imbalance Handling

The compiled dataset (CBDN0_FINAL.csv) exhibits class imbalance with instances predominantly concentrated in the general and gender/sexual harassment categories, while threat-related cases remain severely underrepresented as shown in Figure 2. Such imbalance can bias model learning and degrade minority-class performance. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the TF–IDF (Term Frequency–Inverse Document Frequency) feature space of the training data, enabling continuous interpolation without semantic distortion. The CNN-based sequential branch was trained on the

original text sequences, with random resampling used to preserve alignment with the oversampled TF–IDF (Term Frequency–Inverse Document Frequency) features. This hybrid rebalancing strategy, applied during severity-level training (low, medium, high), improved minority-class recognition and enhanced model generalization, as reflected in Figure 3.

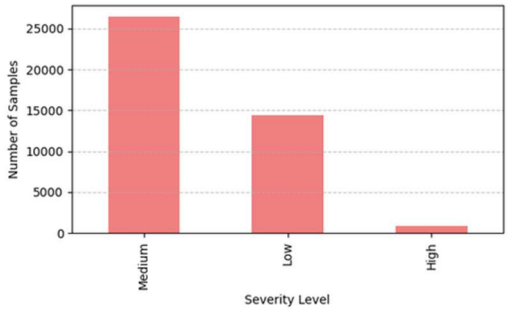

Figure 2: Dataset class distribution before SMOTE

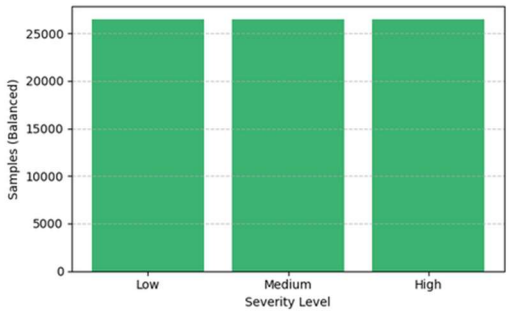

Figure 3: Dataset class distribution after SMOTE

The Synthetic Minority Oversampling Technique (SMOTE) generates a synthetic sample $\tilde{x}$ by interpolating between a minority instance x_i and one of its nearest neighbors x_{nn}, according to the equation (1):

$$\tilde{\mathbf{x}} = \mathbf{x}_i + \lambda(\mathbf{x}_{nn} - \mathbf{x}_i), \quad \lambda \sim U(0,1) \qquad (1)$$

where $\mathbf{x}_i$ represents a minority-class TF-IDF (Term Frequency–Inverse Document Frequency) feature vector, $\mathbf{x}_{nn}$ denotes one of its k-nearest neighbors ($k = 3$ in this study), and λ is a random interpolation coefficient sampled from the continuous uniform distribution $U(0,1)$, which determines the relative position of the synthetic sample along the line segment between x_i and its nearest neighbor x_{nn}. In this work, SMOTE is applied exclusively to the TF-IDF feature space, as this representation supports continuous interpolation without distorting semantic structure. The CNN-based sequential input branch is not synthetically oversampled; instead, original text sequences are randomly resampled to align with the expanded TF-IDF training set, ensuring consistent sample pairing during hybrid model training.

3.4 Model Architecture

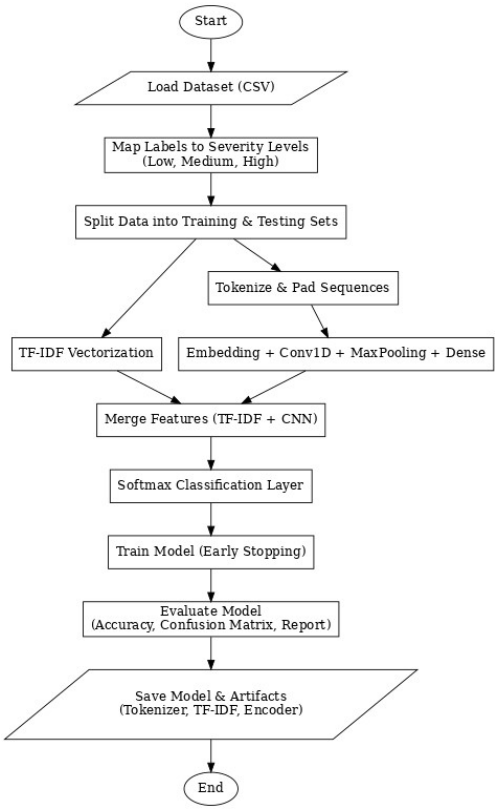

Figure 4: Overall Model Development Workflow

As shown in figure 4, the model development pipeline begins by loading the Malay–English cyberbullying dataset from a CSV file, after which textual labels are mapped to three severity levels: low, medium, and high. The dataset is then partitioned into training and testing subsets, and two parallel feature extraction streams are employed. The TF–IDF (Term Frequency–Inverse Document Frequency) branch transforms text into n-gram–based TF–IDF vectors, while the CNN branch tokenizes and pads the text before passing it through an embedding layer followed by convolutional and pooling operations to capture contextual representations. Outputs from both branches are concatenated and processed through fully connected layers, with a final softmax layer producing severity classifications. The trained model, together with preprocessing artifacts (tokenizer, TF–IDF vectorizer, and label

encoder), is subsequently saved for deployment in the Streamlit application.

3.5 TF-IDF Feature Representation

To quantify lexical importance, the statistical branch adopts the TF-IDF (Term Frequency–Inverse Document Frequency) weighting scheme, a widely established representation in information retrieval and natural language processing. The TF-IDF weight assigned to a term t in a document d is defined as seen in the equation (3).

$$\text{tfidf}(t,d) = \text{tf}(t,d) \cdot \text{idf}(t) \quad (3)$$

where $\text{tfidf}(t,d)$ denotes the final weight of term t in document d,$\text{tf}(t,d)$ represents the term frequency, defined as the number of occurrences of t within d, and $\text{idf}(t)$ corresponds to the inverse document frequency that penalizes terms appearing frequently across the corpus. The inverse document frequency is computed as seen in the equation (2).

$$\text{idf}(t) = \log\left(\frac{N_{\text{train}}+1}{\text{df}(t)+1}\right) + 1 \quad (2)$$

Where N_{train} denotes the total number of documents in the training set, and $\text{df}(t)$ is the document frequency, defined as the number of training documents containing term t. Additive smoothing constants are incorporated to prevent division by zero and to stabilize the weighting of infrequent terms. In this study, unigram and bigram (1-2 gram) features are extracted with a maximum vocabulary size of 2000 terms. The resulting TF-IDF (Term Frequency–Inverse Document Frequency) vectors constitute the inputs to the dense sub-network of the hybrid architecture, providing complementary lexical cues to the CNN-based semantic features.

3.6 Model Training and Testing

Cyberbullying categories were mapped into three severity levels to support graded detection and interventions. High severity was assigned to threat-related content, reflecting explicit or implicit intent to inflict harm, intimidation, or fear, and is widely recognized as the most critical form of online abuse requiring immediate attention [42],[13]. Medium severity encompasses gender/sexual harassment and religious bullying, capturing identity-based abuse that, while not always overtly violent, is associated with sustained psychological harm and significant emotional distress [44],[18]. Low severity includes generalized hostile or derogatory expressions that convey bias or contempt without direct threats, representing early-stage or normalized aggression that may precede escalation [11], [49]. This severity mapping is consistent with established harm-based taxonomies and prior research on graded cyberbullying severity and interventions prioritization, as simplified in the equation (4).

To ensure reproducibility and clarity in the implementation of the proposed hybrid model, the development and training configurations were explicitly defined and standardized prior to experimentation. These settings govern data input, feature representation, model architecture, training strategy, and evaluation protocol. Table 4 summarizes the complete model development and training configuration employed in this study. Following feature extraction, the CNN-derived semantic representation and TF-IDF statistical features are concatenated and passed through fully connected layers.

Table 4: Model Development and Training Configuration

Category	Setting / Description
Dataset Used	*CBDN0_FINAL.csv* – 52,140 text samples (Malay-English cyberbullying corpus)
Input Features	1. Tokenized text sequences (for CNN branch) 2. TF-IDF weighted features (for statistical branch)
Label Mapping	Threat → High Gender/Sexual Harassment → Medium Religious Bullying → Medium General Harassment → Low
Train–Test Split	80% Training / 20% Testing (Stratified by severity class)
Text Preprocessing	Tokenization, padding/truncation to 120 tokens, TF-IDF (1–2 n-gram, max 2000 features), lowercasing
Vectorization	TF-IDF Vectorizer: max_features = 2000, ngram_range = (1,2)

CNN Branch Configuration	Embedding Dimension = 100 Convolution Filters = 128 Kernel Sizes = (3, 4, 5) GlobalMaxPooling1D applied to each Dense(128, ReLU) + Dropout(0.3)
TF-IDF Branch Configuration	Dense(256 → 64, ReLU) + Dropout(0.3)
Fusion Layer	Concatenation of CNN and TF-IDF feature outputs
Output Layer	Dense layer with Softmax activation (3 classes: Low, Medium, High)
Imbalance Handling	SMOTE (Synthetic Minority Oversampling Technique) applied on TF-IDF features only, k_neighbors = 3
Tokenizer Settings	max_words = 20,000; OOV token = <OOV>; max_sequence_length = 120
Training Parameters	Optimizer = Adam Loss = Categorical Crossentropy Epochs = 6 Batch Size = 32 Validation Split = 0.12
Regularization Techniques	SpatialDropout1D (0.2) on embeddings Dropout (0.3) on dense layers Early Stopping (patience = 2)
Random Seed	42 (for reproducibility)
Evaluation Metrics	Accuracy, Loss, Precision, Recall, F1-Score, Confusion Matrix
Model Artifacts Saved	CSD_MODEL_S.h5 (Trained model), tokenizer.pkl, tfidf_vectorizer.pkl, label_encoder.pkl

4. RESULTS AND EVALUATION

4.1 Training and Validation Performance

The model's convergence behavior is presented in figures 5 and 6, which illustrate the training and validation accuracy and loss over six epochs and the results demonstrate rapid learning stability with training accuracy improved consistently from 95% to nearly 99.9%, and the validation accuracy reached approximately 99% by the final epoch, Similarly, both training and validation loss declined sharply, stabilizing below 0.03. The proximity between training and validation curves indicates minimal overfitting, suggesting that regularization techniques such as dropout and early stopping were effective.

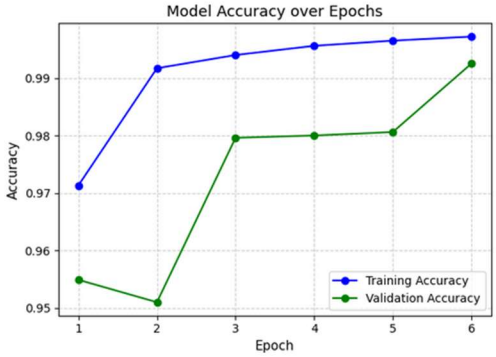

Figure 5: Model Accuracy over Epochs

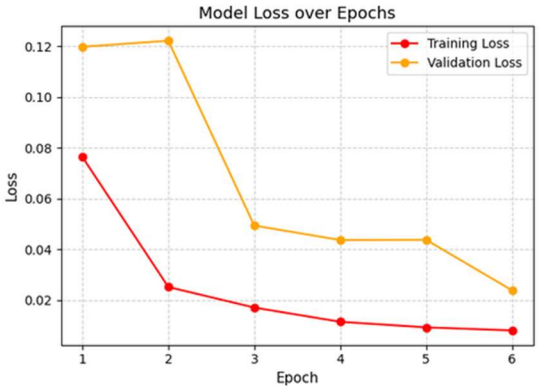

Figure 6: Model Loss over Epochs

4.2 Model Evaluation and Classification Analysis

The model's performance on the test set is summarized in figure 5, showing a test accuracy of 0.9841 and figure 6 showing a test loss of 0.0753, but precision, recall, and F1-scores for each class exceeded 0.93, confirming robust generalization, while the "Medium" and "Low" classes exhibited slightly higher F1-scores (≈0.98), which indicates that the model effectively captured linguistic nuances and context overlap within these categories.

During model training, the CBDN0_FINAL.csv dataset was split into 80% training data and 20% testing data, where the training part was balanced by applying the SMOTE

(Synthetic Minority Oversampling Technique) technique to correct the imbalance between the three severity classes: Low, Medium, and High as shown in figure 3 and 4. The SMOTE was applied only to the training set, because the training set is used for model learning, while the testing set must remain untouched to mimic the real-world application. Even though the training data became balanced after we applied the SMOTE technique, the testing data (20%) remained imbalanced because it came directly from the original dataset and this imbalance appears in the confusion matrix and the classification report. Such behavior is expected because the confusion matrix is generated using the testing data only and the purpose of the confusion matrix is to evaluate how well the model performs on unseen, real-world, balanced or imbalanced data, not on artificially balanced data and if the SMOTE were applied to the testing set, the evaluation would become unrealistic and biased, therefore our confusion matrix reflects the natural imbalance of the 20% testing data while the model itself was trained on a balanced dataset (80%) and this ensures a fair and reliable evaluation of model performance.

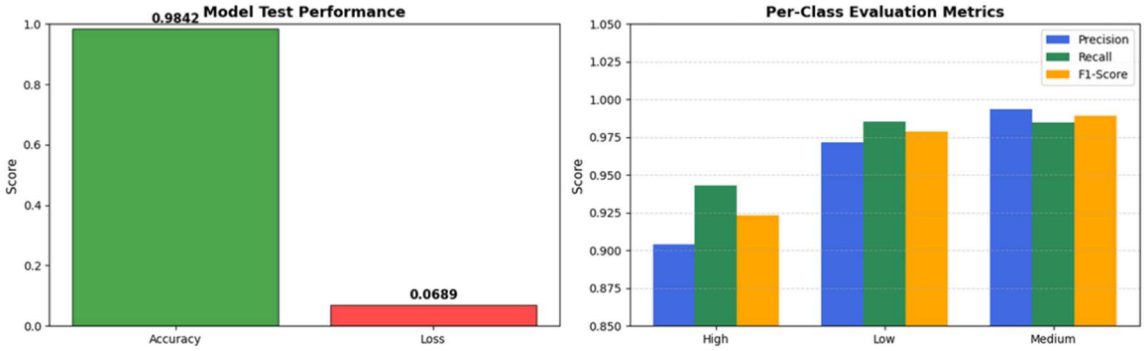

Figure 7: Model Evaluation Summary on the Test Set

The confusion matrix reveals strong predictive accuracy with most instances correctly classified along the diagonal as presented in figure 8. However minor misclassifications occurred between "Low" and "Medium" severity levels, an expected outcome in code-mixed datasets, where tone and lexical choice overlap frequently. Also, the macro and weighted averages (precision = 0.9844, recall = 0.9841, F1 = 0.9842) indicate balanced performance across all categories. The results demonstrate that combining convolutional semantic features with TF-IDF statistical representations significantly improves detection of cyberbullying severity, where the CNN captures phrase-level sentiment and aggression expressions, while TF-IDF of NLP contributes lexical weighting, enabling precise classification even with code-switching between Malay and English. The use of SMOTE effectively countered data imbalance especially for the minority "High" class, aligning with prior findings that synthetic oversampling enhances minority representation without degrading model precision.

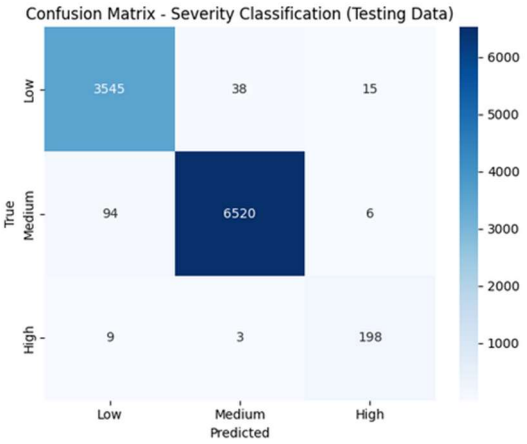

Figure 8: Confusion Matrix for Severity Classification on the Test Set

4.3 Model Ablation Study

To quantify the contribution of individual architectural components, an ablation study was conducted by isolating the sequence-based CNN branch and the lexical TF-IDF branch while keeping all other settings unchanged. Table 5 reports the ablation study results of the proposed hybrid architecture.

Table 5: Ablation Study of the Proposed Hybrid Architecture

Model Variant	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
Hybrid CNN + TF-IDF (SMOTE)	0.9848	0.9707	0.9635	0.9670	0.9848
TF-IDF-only (MLP)	0.9841	0.9713	0.9649	0.9680	0.9841
CNN-only (Sequence)	0.0201	0.0067	0.3333	0.0132	0.0008

In the ablation test, the CNN-only model showed weak performance (accuracy = 2.01% and macro F1 = 0.0132), suggesting the sequence-based contextual modeling alone is insufficient for severity classification of Malay-English code-mixed cyberbullying text. The TF-IDF-only model produced strong performance (macro F1 = 0.9680), further supporting that lexical cues are important for identifying abusive content; yet this approach remains limited in capturing deeper contextual and semantic dependencies. Combining these representations, a hybrid CNN + TF-IDF model provides the most balanced and robust performance across all severity classes.

4.4 Model Baseline Comparison

To further contextualize performance, the proposed hybrid model was compared against a classical Logistic Regression baseline trained on TF-IDF features**.** Table 6 compares the proposed hybrid CNN + TF-IDF model with a classical TF-IDF–based Logistic Regression baseline, highlighting the benefit of combining contextual and lexical features for more balanced severity classification.

Table 6: Comparison with Classical Baseline Classifiers

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
Hybrid CNN + TF-IDF (SMOTE)	0.9848	0.9707	0.9635	0.9670	0.9848
Logistic Regression + TF-IDF	0.9817	0.9442	0.9767	0.9596	0.9818

While the Logistic Regression baseline achieves competitive recall, it underperforms the hybrid model in macro-averaged metrics. The proposed architecture yields a higher macro F1-score (0.9670 vs. 0.9596), indicating improved balance across severity classes, including minority high-severity cases. These results demonstrate that integrating neural contextual modeling with lexical representations provides measurable gains beyond linear decision boundaries.

4.5 User Behavior Profiling

Following the severity classification phase, a rule-based user behavior profiling framework was developed to complement the hybrid NLP-CNN model and enhance its interpretability. The profiling module provides qualitative insight into user interaction tendencies, and this additional layer enables the model to infer behavioral context from the linguistic composition of posts, offering a more nuanced understanding of online aggression patterns and by integrating linguistic and behavioral profiling, the system strengthens transparency, allowing researchers, moderators, and policymakers to interpret automated predictions through human-centered reasoning, and moreover this approach aligns with contemporary digital safety research emphasizing the importance of connecting textual aggression indicators with user intent for adaptive interventions design [39], [13], [9].

The profiling mechanism is based on keyword-conditioned heuristic rules grounded in linguistic and socio-psychological studies of online communication. These heuristics associate predicted severity levels with lexical, tonal, and contextual cues that reflect underlying user motivation. Prior empirical work demonstrates that markers such as sarcasm, humor, exaggeration, and explicit hostility are reliable indicators of aggression orientation and emotional impulsivity [42], [44]. Building on this evidence, the framework infers behavioral archetypes: provocative, impulsive, harassing, and defensive, by identifying recurring linguistic patterns consistent with these psychological profiles. Table 7 outlines the rule-based user behavior profiling framework, mapping predicted severity levels and linguistic cues to distinct behavioral categories grounded in established theoretical and empirical studies.

Table 7: Rule-Based User Behavior Profiling Framework

Behavior Profile	Rule Logic	Indicative Keywords	Interpretation And Theoretical Basis	References
Impulsive	If *severity = Low* and text contains any humor or casual intent markers, OR if no offensive keyword is detected, classify as *Impulsive*.	"joke", "main-main", "saja", "lol", "haha"	Represents *situational teasing or emotional outburst* rather than persistent hostility. Reflects impulsive online expression typical among adolescents and young users.	[39][13]
Provocative / Impulsive	If *severity = Medium* and text contains direct insults or identity-based terms, classify as *Provocative*; otherwise, retain *Impulsive*.	"stupid", "useless", "no brain", "bodoh", "bangang"	Indicates *intentional provocation* aimed at social dominance or peer ridicule but without explicit violent threat. Balances verbal aggression with cultural elements of sarcasm and humor.	[44][18]
Harassing	If *severity = High* and text contains violent verbs or coercive expressions, classify as *Harassing*.	"kill", "hurt", "mati", "bunuh", "sampah"	Denotes *explicit aggression* and intent to cause harm. Represents the most severe behavioral state within online hostility typologies.	[48][42]
Defensive	If *severity = High* and text includes self-protective or resistant expressions, classify as *Defensive*.	"stop", "leave me", "diam lah", "cukup", "back off"	Signals *resistance or retaliatory defense* rather than aggression. Reflects reactive behavior to perceived harassment or attack.	[9][39]

4.6 Application System Testing

To evaluate the practical applicability of the proposed hybrid NLP-CNN model, a Streamlit-based web application was developed for real-time cyberbullying detection and user behavior profiling in Malay-English text and the application serves as an interactive demonstration of the model's operational workflow, providing an accessible interface for both researchers and system moderators, users can input any short social media post or comment written in code-mixed Malay-English, upon which the system executes two parallel analyses: (i) *Severity Level Classification* by assigning each input to one of three severity levels (*Low*, *Medium*, or *High*) based on the trained hybrid model, and (ii) *User Behavior Profiling* by inferring behavioral patterns such as *Impulsive*, *Provocative*, *Defensive*, or *Harassing* through rule-based keyword mapping. This integration bridges computational detection with linguistic interpretation, advancing the interpretability of automated moderation systems [50], [51].

4.7 Model Confidence Score

To improve transparency and practical usability, the proposed hybrid NLP-CNN model reports confidence scores together with each predicted severity label. Let x denote the processed input text, formed by combining the CNN-based sequence representation x_{seq} and the TF-IDF feature vector x_{tfidf}. After feature fusion, the network produces a real-valued logit vector $z = (z_1, z_2, z_3)$, where each $z_k \in \mathbb{R}$ represents the un-scaled evidence for class k (with $k = 1,2,3$ corresponding to low, medium, and high severity). These values are transformed into probabilities using the softmax function of the equation (4).

$$P(y = k \mid \mathbf{x}) = \frac{\exp(z_k)}{\sum_{j=1}^{3} \exp(z_j)}, \quad (4)$$

Where $P(y = k \mid \mathbf{x})$ denotes the posterior probability of a class k, $\exp(\cdot)$ is the exponential operator, and the denominator ensures that the probabilities sum to one. The predicted class is defined as $\hat{y} = \text{argmax}_k P(y = k \mid \mathbf{x})$, and the quantity $P(y = \hat{y} \mid \mathbf{x})$ is taken as the model's confidence. In practice, higher probabilities indicate clearer lexical and contextual signals of severity, while more evenly distributed values suggest ambiguous or overlapping expressions, which frequently occur in Malay-English code-mixed content. Presenting these scores provides a simple measure of reliability, allowing high-confidence high-risk cases to be prioritized and uncertain predictions to be reviewed manually, thereby supporting more accountable and informed moderation decisions.

Figure 9 presents representative system-level testing of the proposed framework using diverse Malay–English cyberbullying prompts. For each test case, the first panel displays the input text together

with the predicted severity level and corresponding behavior profile, while the second panel reports the class-wise confidence scores produced by the model.

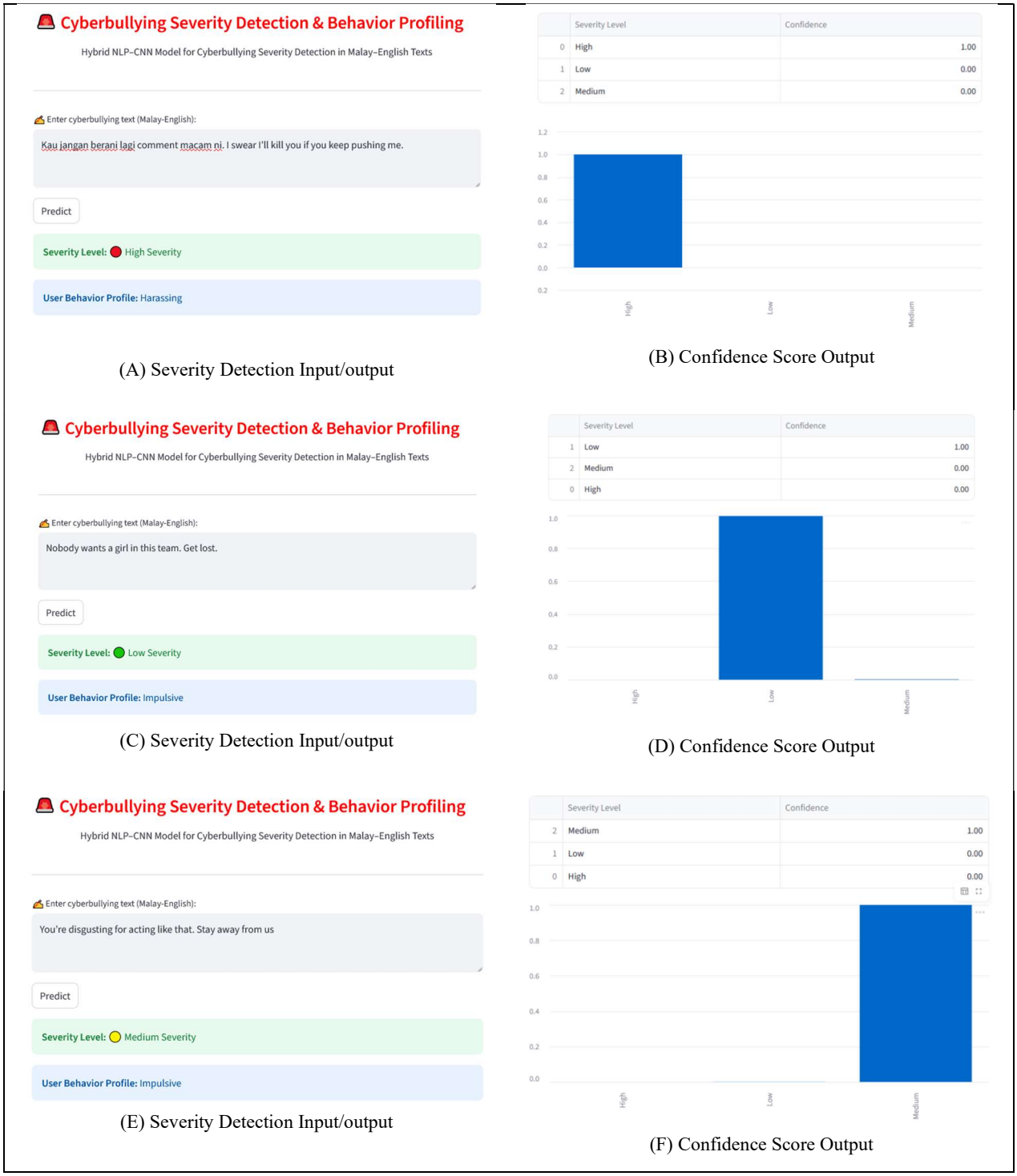

Figure 9: Streamlit-Based Demonstration of the Proposed Cyberbullying Severity Detection and User Behavior Profiling

Figure 9 illustrates the system interface and representative test outputs of the proposed cyberbullying detection framework. Panel (A) displays a sample input of Malay-English cyberbullying text, along with the predicted severity level classified as High and the corresponding user behavior profile labeled as Harassing. Panel (B) shows the associated confidence scores for each

severity class: Low (0.00), Medium (0.00), and High (1.00). In this instance, the model accurately identifies the text as high severity and assigns the appropriate behavioral profile according to the predefined severity taxonomy and profiling framework.

The confidence scores, derived from the softmax layer of the hybrid NLP-CNN model, represent the model's probabilistic support for each class. Rather than producing a single deterministic output, the model distributes probability across Low, Medium, and High severity categories. Higher confidence values indicate clearer linguistic evidence, while lower scores typically correspond to texts containing ambiguous or overlapping expressions. Panels (C–F) follow the same representation, illustrating additional examples of severity classification and behavioral profiling. A comprehensive summary of all test cases is provided in table 8.

Table 8: System Test Logs

Test Case	Text Input	Severity & User Profile Output	Confidence Score (Low / Medium / High) Output
1	*Kau jangan berani lagi comment macam ni. I swear I'll kill you if you keep pushing me.*	High / Harassing	0.00 / 0.00 / 1.00
2	*Nobody wants a girl in this team. Get lost*	Low / Impulsive	1.00 / 0.00 / 0.00
3	*You're disgusting for acting like that. Stay away from us*	Medium / Impulsive	0.00 / 1.00 / 0.00
4	*Tak payah over sangat la, you're always so dramatic and annoying It's just a stupid game je. Relax lah, don't embarrass yourself.*	Medium / Impulsive	0.00 / 1.00 / 0.00
5	*Seriously ah, kau ni memang tak guna. Just delete your account la*	Medium / Harassing	0.00 / 1.00 / 0.00
6	*Next time aku jumpa kau, kau akan menyesal hidup kau.*	High / Harassing	0.00 / 0.00 / 1.00
7	*Berani cakap macam tu? Satu tikaman je cukup untuk kau*	High / Harassing	0.00 / 0.00 / 1.00
8	*Orang macam kau selalu jumpa nasib buruk sendiri*	Low / Provocative	1.00 / 0.00 / 0.00
9	*Alah, kau masih ikut puasa old school la*	Medium / Provocative	0.00 / 1.00 / 0.00
10	*Alah, kau lambat je selalu… buat malu group je.*	Low / Impulsive	1.00 / 0.00 / 0.00

The proposed system exhibits strong performance in identifying the severity of cyberbullying instances and profiling user behavior within code-mixed Malay-English texts. High-confidence predictions correspond to clearly identifiable linguistic indicators of harassment or impulsive behavior, while medium-confidence outputs reveal subtler, context-dependent expressions that challenge conventional detection methods. The integration of natural language processing with CNN-based feature extraction enables a probabilistic and interpretable framework, allowing the system to quantify uncertainty while maintaining classification accuracy.

These findings highlight the advantages of hybrid NLP-CNN approach in capturing both explicit and nuanced cyberbullying patterns, supporting scalable, context-sensitive interventions. By providing actionable insights into user behaviour and message severity, the model not only advances automated detection capabilities but also informs the design of preventive and responsive strategies for online safety. Overall, the results underscore the feasibility and effectiveness of leveraging CNN-NLP approaches for nuanced, real-world cyberbullying detection in multilingual text data.

5. CONCLUSION

This study demonstrates that cyberbullying severity detection in Malay-English code-mixed text can be addressed effectively through a hybrid NLP-CNN framework that integrates statistical lexical features with deep semantic representations. The proposed model offers a scalable and adaptable computational framework that advances severity-aware abuse detection and supports reproducible research, real-world moderation, and evidence-based policymaking in code-switching prevalent environments such as Malay-English code-mixed. Furthermore, beyond predictive performance, user behavior profiling adds interpretive value by relating severity scores to behavioral tendencies such as impulsive, provocative, defensive, and harassing actions. This extra layer of transparency supports human-centered moderation and facilitates more context-aware intervention models. Further work can advance this framework towards broader cross-platform datasets, context-dependent embedding models, or multimodal inputs to add to this work, offering potentially even greater effectiveness in terms of proactive and responsible mitigation of cyberbullying in the wide variety of digital environments.

5.1 Comparison with Prior Approaches

Compared to prior studies, which predominantly address binary detection or severity classification in monolingual settings, the proposed framework operates within a more complex Malay–English code-mixed context while modeling cyberbullying across graded severity levels. Earlier approaches—ranging from rule-based systems to classical machine learning and single-branch deep learning models—primarily rely on either lexical cues or contextual embeddings, limiting their ability to capture the combined linguistic and semantic variability inherent in code-mixed discourse. In contrast, the proposed hybrid NLP–CNN architecture integrates TF-IDF and convolutional representations, enabling complementary learning of surface-level and contextual features.

Empirically, while previous studies report strong performance under controlled monolingual conditions, the proposed model achieves 98.42% accuracy with consistently high precision, recall, and F1-scores across all severity classes in a multilingual setting. The observed gains are further supported by ablation and baseline comparisons, which show that the hybrid configuration yields more balanced performance than single-representation models and classical classifiers, particularly for minority high-severity instances. In addition, the integration of user behavior profiling extends beyond conventional severity classification, providing an interpretable layer that links predictive outcomes to interaction patterns, a dimension largely absent in earlier work.

5.2 Implication of the Study

This study demonstrates that cyberbullying severity detection in Malay–English code-mixed discourse is both technically feasible and methodologically robust when hybrid feature integration is employed. The superior and balanced performance of the proposed NLP–CNN architecture confirms that combining lexical representations with contextual semantic modelling is essential for capturing graded harm in multilingual and informal online communication, where single-representation models are empirically insufficient.

The construction and standardization of a large-scale severity-annotated Malay–English dataset directly address a critical resource gap in cyberbullying research. By operationalizing severity into low, medium, and high levels grounded in harm-based categorization, this study provides a replicable framework for severity modelling in low-resource and code-switched settings, enabling reproducibility and comparative benchmarking that have been largely absent in prior work.

Practically, the shift from binary detection to severity-aware classification enables proportionate and risk-sensitive intervention. Differentiating between mild harassment and high-risk threats supports more targeted moderation, safeguarding, and prevention strategies, reducing both under-reaction to severe abuse and over-penalization of low-severity cases. The integration of rule-based user behaviour profiling further strengthens interpretability by contextualizing model outputs in terms of observable behavioural tendencies rather than opaque predictions. At the system level, the model's stable performance across severity classes and its confidence-based output demonstrate suitability for real-time deployment in digital safety applications. This reliability is particularly significant for high-severity cases, where misclassification carries ethical and practical consequences.

More broadly, the findings reinforce severity as a central analytical dimension in cyberbullying research rather than an auxiliary extension of abuse detection. For multilingual societies characterized by pervasive code-switching, severity-aware and linguistically grounded models offer a more accurate

and culturally aligned basis for digital harm assessment, policy formulation, and responsible AI-mediated moderation.

5.3 Limitation of the Study

The observed limitations are primarily associated with the evaluation scope and the granularity of modeling, rather than the core design of the proposed framework. The model operates at the level of individual text instances, which, while effective for localized severity detection, does not explicitly account for sequential interaction patterns such as repetition, escalation, or sustained targeting that typically unfold across conversation threads. In addition, performance analysis indicates that classification sensitivity is most challenged at adjacent severity boundaries, where overlapping linguistic expressions in Malay–English code-mixed discourse introduce subtle ambiguity. This reflects the intrinsic complexity of distinguishing closely related levels of harmful intent, rather than a structural limitation of the model itself.

5.4 Recommendations for Future Research

Based on the findings of this study, future research should extend the proposed framework beyond single-message severity classification toward interaction-level analysis. While the hybrid NLP–CNN model effectively identified cyberbullying severity using lexical, semantic, and behavioural features, the current approach evaluates messages independently. Future work should therefore examine sequences of user interactions within conversation threads to better understand how cyberbullying behaviour develops and escalates over time, allowing for earlier and more informed intervention decisions.

Future studies should also prioritize the development of larger and standardized Malay–English severity-annotated datasets to improve reproducibility and enable meaningful benchmarking across studies. Expanding dataset coverage to include a wider range of online platforms and variations of Malay–English code-mixed communication would strengthen model generalization and support more reliable evaluation of severity detection systems in multilingual online environments.

ACKNOWLEDGMENT

The authors are grateful to Universiti Tenaga Nasional (UNITEN) for the academic support that made this work possible. Ibrahim Inusa appreciates and extends his thanks to the supervisors and collaborators whose guidance and contributions helped shape the development of this research.

REFERENCES

[1] Jimerson, S. R., Swearer, S. M., & Espelage, D. L. (Eds.). (2010). *Handbook of bullying in schools: An international perspective*. Routledge.

[2] Moreno-Ruiz, D., Martínez-Ferrer, B., & García-Bacete, F. (2019). Parenting styles, cyberaggression, and cybervictimization among adolescents. *Computers in Human Behavior, 93*, 252–259. https://doi.org/10.1016/j.chb.2018.12.031

[3] Sathyanarayana Rao, T. S., Bansal, D., & Chandran, S. (2018). Cyberbullying: A virtual offense with real consequences. *Indian journal of psychiatry*, *60*(1), 3–5. https://doi.org/10.4103/psychiatry.IndianJPsychiatry_147_18

[4] Kowalski, R. M., & Limber, S. P. (2007). Electronic bullying among middle school students. *Journal of Adolescent Health, 41*(6, Suppl.), S22–S30. https://doi.org/10.1016/j.jadohealth.2007.08.017

[5] Saruddin, M. Z., Azman, A. Z. F., Zulkefli, N. A. M., & Isa, M. R. (2025). Prevalence of cyberbullying and its associated factors among high school students in Klang District, Malaysia. *Malaysian Journal of Medicine and Health Sciences, 21*(3), 220–231. https://doi.org/10.47836/mjmhs.21.3.26

[6] Teng, T.H., Varathan, K.D., & Crestani, F. (2023). A comprehensive review of cyberbullying-related content classification in online social media. Expert Syst. Appl., 244, 122644.

[7] Perera, A., & Fernando, P. (2024). Cyberbullying detection system on social media using supervised machine learning. *Procedia Computer Science, 239*, 506–516. https://doi.org/10.1016/j.procs.2024.06.200

[8] Maity, K., Jha, P., Jain, R., Saha, S., & Bhattacharyya, P. (2024). *"Explain thyself bully": Sentiment aided cyberbullying detection with explanation* (arXiv:2401.09023). arXiv. https://arxiv.org/abs/2401.09023

[9] Wu J-L, Tang C-Y. Classifying the Severity of Cyberbullying Incidents by Using a Hierarchical Squashing-Attention Network.

Applied Sciences. 2022; 12(7):3502. https://doi.org/10.3390/app12073502

[10] Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-assisted content analysis: Using large language models to support deductive coding (arXiv:2306.14924). arXiv. https://arxiv.org/abs/2306.14924

[11] Talpur BA, O'Sullivan D (2020) Cyberbullying severity detection: A machine learning approach. PLoS ONE 15(10): e0240924. https://doi.org/10.1371/journal.pone.0240924

[12] Sahlan, F., Hamidi, F., Zulhafizal Misrat, M., Adli, M. H., Wani, S., & Gulzar, Y. (2021). Prediction of Mental Health Among University Students. *International Journal on Perceptive and Cognitive Computing*, *7*(1), 85–91. Retrieved from https://journals.iium.edu.my/kict/index.php/IJPCC/article/view/225

[13] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073–1137. https://doi.org/10.1037/a0035618

[14] Mahmud, S., Ali, A., & Khusro, M. (2023). Multilingual cyberbullying detection in social media: A Malaysian perspective. *International Journal of Information Management, 63*, 101359. https://doi.org/10.1016/j.ijinfomgt.2022.101359

[15] Pawar, R., & Raje, R.R. (2019). Multilingual Cyberbullying Detection System. 2019 IEEE International Conference on Electro Information Technology (EIT), 040-044

[16] Md Nasir, A. F., & Mohamad Sukri, K. A. (2022). Machine learning approach on cyberstalking detection in social media using Naive Bayes and decision tree. *Journal of Soft Computing and Data Mining, 3*(1), 19–27. https://doi.org/10.30880/jscdm.2022.03.01.002

[17] Monirah Abdullah Al-Ajlan and Mourad Ykhlef. "Deep Learning Algorithm for Cyberbullying Detection". International Journal of Advanced Computer Science and Applications (IJACSA) 9.9 (2018). http://dx.doi.org/10.14569/IJACSA.2018.090927

[18] Vyawahare, M. ., & Govilkar, S. . (2022). Identifying Severity of Cyberbullying Using Scalable Labeled Multi-Platform Dataset. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(4), 201–210. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2217

[19] Prama, Tabia Tanzin and Amrin, Jannatul Ferdaws and Anwar, Md Musfique and Sarker, Iqbal H., Ai-Enabled User-Specific Cyberbullying Severity Detection with Explainability. Available at SSRN: https://ssrn.com/abstract=5167003 or http://dx.doi.org/10.2139/ssrn.5167003

[20] Obaid, M.H., Guirguis, S.K., & Elkaffas, S.M. (2023). Cyberbullying Detection and Severity Determination Model. IEEE Access, 11, 97391-97399.

[21] Johari, N. F. B., & Jaafar, J. (2022). A Malay language cyberbullying detection model on Twitter using supervised machine learning. In *2022 International Visualization, Informatics and Technology Conference (IVIT)* (pp. 325–332). IEEE. https://doi.org/10.1109/IVIT55443.2022.10033395

[22] Unnava, S., & Parasana, S. R. (2024). A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach. *Engineering, Technology & Applied Science Research*, *14*(4), 15607–15613. https://doi.org/10.48084/etasr.7621

[23] Rahman-Laskar, S., Gupta, G., Badhani, R., & Pinto-Avendaño, D. E. (2024). Cyberbullying detection in a multi-classification codemixed dataset. *Computación y Sistemas, 28*(3), 1091–1113. https://doi.org/10.13053/CyS-28-3-4989

[24] Rezvan, M., Shetty, S., & Mather, P. (2018). A dataset for harassment research. In *ASONAM 2018* (pp. 1385–1388). IEEE.

[25] Ogunleye, B., & Dharmaraj, B. (2023). The Use of a Large Language Model for Cyberbullying Detection. *Analytics*, *2*(3), 694-707. https://doi.org/10.3390/analytics2030038

[26] Alotaibia, A. F., AlZain, M. A., Masud, M., & Jhanjhi, N. Z. (2021). A comprehensive survey on security threats and countermeasures of cloud computing environment. *Turkish Journal of Computer and Mathematics Education, 12*(9), 1978–1990.

[27] Qian, Y., Zhang, W., & Liu, T. (2023). Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational*

Linguistics: EMNLP 2023 (pp. 6516–6528). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.433

[28] Palladino, B. E., Menesini, E., Nocentini, A., Luik, P., Naruskov, K., Ucanok, Z., Dogan, A., Schultze-Krumbholz, A., Hess, M., & Scheithauer, H. (2017). Perceived severity of cyberbullying: Differences and similarities across four countries. *Frontiers in Psychology, 8*, 1524. https://doi.org/10.3389/fpsyg.2017.01524

[29] Koehler, C., & Weber, M. (2018). Do I really need to help?!" Perceived severity of cyberbullying, victim blaming, and bystanders' willingness to help the victim. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *12*(4), Article 4. https://doi.org/10.5817/CP2018-4-4

[30] Makkala, N., Plubin, B., Bunyatisai, W., Mouktonglang, T., & Plubin, S. (2025). Classifying the severity of cyberbully from social media comments. *International Journal of Computer Applications, 186*(63), 34–42. https://doi.org/10.5120/ijca2025924448

[31] Hollá, K., Fenyvesiová, L., & Hanuliaková, J. (2017). Measurement of cyber-bullying severity. *The New Educational Review.* https://doi.org/10.15804/tner.2017.47.1.02

[32] Huang, L., Li, W., Xu, Z., Sun, H., Ai, D., Hu, Y., Wang, S., Li, Y., & Zhou, Y. (2023). The severity of cyberbullying affects bystander intervention among college students: The roles of feelings of responsibility and empathy. *Psychology Research and Behavior Management, 16,* 893–903. https://doi.org/10.2147/PRBM.S397770

[33] Aggarwal, A., Maurya, K., & Chaudhary, A. (2020). Comparative Study for Predicting the Severity of Cyberbullying Across Multiple Social Media Platforms. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 871-877.

[34] Rahman-Laskar, S., Gupta, G., Badhani, R., & Pinto-Avendaño, D. E. (2024). Cyberbullying detection in a multi-classification codemixed dataset. *Computación y Sistemas, 28*(3), 1091–1113. https://doi.org/10.13053/CyS-28-3-4989

[35] Ahmad, Z., Maskat, R., & Mohamed, A. (2023). Harnessing natural language processing for mental health detection in Malay text: A review. In *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (pp. 29–35). IEEE. https://doi.org/10.1109/AiDAS60501.2023.10284653

[36] Ahmadinejad, M., Shahriar, N., & Fan, L. (2023). Self-training for Cyberbullying detection: Achieving high accuracy with a balanced multi-class dataset. *Proceedings*. https://api.semanticscholar.org/CorpusID:263734457

[37] Maity, K., Bhattacharya, S., Saha, S., & Seera, M. (2023). A deep learning framework for the detection of Malay hate speech. *IEEE Access*, *11*, 79542-79552. https://doi.org/10.1109/ACCESS.2023.3298808

[38] Maskat, R., Azman, N. A., Nulizairos, N. S. S., Zahidin, N. A., Mahadi, A. H., Norshamsul, S. R., Sharif, M. M. M., & Mahdin, H. (2024). A bi-annotated Malay-English code-switching (Manglish) dataset of X posts for biological gender identification and authorship attribution. *Data in brief*, *52*, 110034. https://doi.org/10.1016/j.dib.2024.110034

[39] Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A., & Parti, K. (2017). Defining Cyberbullying. *Pediatrics*, *140*(Suppl 2), S148–S151. https://doi.org/10.1542/peds.2016-1758U

[40] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

[41] Loh, X.-K., Lee, V.-H., Loh, X.-M., Tan, G. W.-H., Ooi, K.-B., & Dwivedi, Y. K. (2021). The dark side of mobile learning via social media: How bad can it get? *Information Systems Frontiers, 24*(6), 1887–1904. https://doi.org/10.1007/s10796-021-10202-z

[42] Barlett, C., & Coyne, S. M. (2014). A meta-analysis of sex differences in cyber-bullying behavior: the moderating role of age. *Aggressive behavior*, *40*(5), 474–488. https://doi.org/10.1002/ab.21555

[43] Vandebosch, H., & Van Cleemput, K. (2008). Defining Cyberbullying: A Qualitative Research into the Perceptions of Youngsters. CyberPsychology & Behavior, 11, 499-503. https://doi.org/10.1089/cpb.2007.0042

[44] Sood, S. M. M., Hua, T. K., & Hamid, B. A. (2020). Cyberbullying through intellect-related insults. *Jurnal Komunikasi: Malaysian Journal of Communication*. Universiti Kebangsaan

Malaysia Press. http://hdl.handle.net/123456789/470

[45] Ali, W. N. H. W., Fauzi, F., & Mohd, M. (2021). Identification of Profane Words in Cyberbullying Incidents within Social Networks. *Journal of Information Science Theory and Practice*, *9*(1), 24–34. https://doi.org/10.1633/JISTAP.2021.9.1.2

[46] Sazali, H., Matondang, R., Mukhtar, D. Y., & Rasyidah, R. (2024). Virtual violence and New Media ethics: The boundary between legitimate and harmful expressions. *Jurnal Kawistara, 14*(3). https://doi.org/10.22146/kawistara.98627

[47] Peebles E. (2014). Cyberbullying: Hiding behind the screen. *Paediatrics & child health*, *19*(10), 527–528. https://doi.org/10.1093/pch/19.10.527

[48] Drake, B. (2015, June 1). *The darkest side of online harassment: Menacing behavior*. Pew Research Center. https://www.pewresearch.org/short-reads/2015/06/01/the-darkest-side-of-online-harassment-menacing-behavior/

[49] Seraj, Z. (2024, July 19). *What counts as cyberbullying? Experts list six ways victims could be harassed online*. Malay Mail. https://www.malaymail.com/news/malaysia/2024/07/19/what-counts-as-cyberbullying-experts-list-six-ways-victims-could-be-harassed-online/144156

[50] Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

[51] Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys, 51*(4), Article 85. https://doi.org/10.1145/3232676

[52] Sadiq, R. B., Safie, N., Abd Rahman, A. H., & Goudarzi, S. (2021). Artificial intelligence maturity model: a systematic literature review. *PeerJ. Computer science*, *7*, e661. https://doi.org/10.7717/peerj-cs.661

[53] Munirah 'Izzati Ismail, and Nor Eleyana Abdullah, (2023) *An Analysis of code-switching in Hitz FM instagram comments.* e-Bangi Journal of Social Sciences and Humanities, 20 (4). pp. 154-163. ISSN 1823-884x

[54] Prama, T.T., Amrin, J.F., Anwar, M.M., & Sarker, I.H. (2025). AI Enabled User-Specific Cyberbullying Severity Detection with Explainability. ArXiv, abs/2503.10650.

[55] Yusop, N., & Al-Shami, S. A. (2021). Risk and protecting factors of cyberbullying in Malaysia: A comparative analysis. *'Ulum Islamiyyah: The Malaysian Journal of Islamic Sciences, 33*(S5), 101–112. https://doi.org/10.33102/uij.vol33noS5.406

[56] Singh, S., & Othman, S. H. (2025). An effective cyberbullying detection model for the Malay language using transformer model on social media platform X. *International Journal of Innovative Computing, 15*(1), 63–71. https://doi.org/10.11113/ijic.v15n1.520

[57] Guo, X., Adnan, H. M., & Abidin, M. Z. Z. (2024). Detecting offensive language on Malay social media: A zero-shot, cross-language transfer approach using dual-branch mBERT. *Applied Sciences, 14*(13), 5777. https://doi.org/10.3390/app14135777

[58] Farhana, A., Shamala, P., & Md Jais, I. (2020). Malaysia-English code-mixing insertion: Why "lepaking" in preference to "hanging out"? *Quantum Journal of Social Sciences and Humanities, 1*(5), 69–84.

[59] Zuwardi, N. A. I., & Razali, F. N. (2024). Code-mixing and code-switching language among Malay language learners in private universities. *Asian Journal of Research in Education and Social Sciences, 6*(7), 8–

[60] Treffers-Daller, J., Majid, S., Thai, Y. N., & Flynn, N. (2022). Explaining the diversity in Malay-English code-switching patterns: The contribution of typological similarity and bilingual optimization strategies. *Languages, 7*(4), 299. https://doi.org/10.3390/languages7040299

[61] Rahim, N. U. A., & Mustapha, N. (2024). Transformer-based model with CNN and CapsNets to improve Malay hate speech detection in tweets. *Journal of Theoretical and Applied Information Technology, 102*(19), 7091–7102.