



ARTICLE

Enhancing Cyberbullying Detection in Social Media through Attention-Based Deep Neural Networks

Mohammad J. M. Zedan^{1,2}, Edwin Rangga Ardhana¹, Siti Raihanah Abdani³
and Mohd Asyraf Zulkifley^{1,*}

¹Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia

²Computer and Information Engineering Department, College of Electronics Engineering, Ninevah University, Mosul, Iraq

³School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia

*Corresponding Author: Mohd Asyraf Zulkifley. Email: asyraf.zulkifley@ukm.edu.my

Received: 26 June 2026; Accepted: 04 September 2026; Published: 22 October 2026

ABSTRACT: The widespread usage of social media has contributed to the global concern of a rapid rise in cyberbullying. Generally, cyberbullying can be defined as intentional and harmful behavior conducted through online communication channels that can cause severe psychological distress, depression, and emotional trauma among victims. Detecting cyberbullying content automatically remains a challenging task because abusive messages are frequently expressed in subtle and ambiguous forms. Consequently, there is an increasing need for intelligent detection systems that are capable of identifying cyberbullying incidents accurately and efficiently. Recent advances in deep learning have demonstrated considerable potential for cyberbullying detection. However, many conventional deep learning models process textual information uniformly, without emphasizing any critical word or phrase that can be a strong indicator of abusive intent. As a result, important features may have been overlooked, which leads to suboptimal detection performance. To address these limitations, this study proposes an enhanced cyberbullying detection framework based on a lightweight attention-enhanced convolutional neural network architecture. This proposed framework employs a compact, one-dimensional neural network architecture to efficiently extract embedded linguistic patterns from short tweets, which are typically extracted using highly complex recurrent networks or transformers. The proposed model was systematically optimized through extensive hyperparameter tuning to identify the most effective configuration of lightweight network layers. In addition, an advanced attention mechanism is integrated to enable the model to focus selectively on bullying-related expressions and contextual cues. This resulted in an improved design capable of detecting most categories of cyberbullying while maintaining low computational complexity and low inference time, which makes it suitable for practical deployment in resource-constrained environments. The proposed model was trained and evaluated using a publicly available cyberbullying dataset containing tweets categorized into five bullying target classes: religion, ethnicity, gender, age, and other cyberbullying, together with an additional non-cyberbullying class. Experimental results demonstrate that the proposed model achieves an accuracy of 83.11% and an F1-score of 82.17%. Furthermore, it also reduces model parameters and floating-point operations by 30.8% and 71.7%, respectively, while maintaining a low inference latency of only 2.5349 ms per sample.

KEYWORDS: Cyberbully identification; text processing; deep learning; social media; attention-based network

1 Introduction

Social media platforms of all kinds have exacerbated the spread of cyberbullying, making it a major global concern. Cyberbullying can have harmful effects, causing psychological suffering for victims such as anxiety and depression, leading to a decline in quality of life and, in extreme cases, even suicide [1–4]. Consequently, the development of advanced automated systems for detecting cyberbullying has gained increasing importance to protect vulnerable groups and prosecute perpetrators. Developing this type of detection system remains a challenging task, as the content of textual posts is often ambiguous or contains complex messages with sarcasm and implicit meanings. In addition, some platforms like Twitter have some posts with 280 characters, making it difficult to interpret such content [5–7]. In recent years, deep learning techniques have shown promising performance in text-based classification tasks, particularly in the context of social media, where convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been effectively applied in detecting cyberbullying [8–10]. However, many existing deep learning detection approaches rely on memory-dependent structures to recognize contextual word association [11,12]. In general, such models are highly effective at understanding the context of written texts, but they remain incapable of interpreting short, word-limited phrases in social media posts that may contain bullying content. Accordingly, deep learning models have been supported by incorporating various types of attention mechanisms to selectively focus on keywords with significant content and prominent textual features associated with abusive behavior [13]. Despite the progress made, there is a substantial need for improvement and room for developing more effective, low-computational models for classifying text posts on various social media platforms. In this work, a one-dimensional CNN integrated with an attention mechanism is investigated to detect cyberbullying in tweet content. Repeated experiments were conducted on a public dataset of approximately 47,000 tweets that were labelled into five categories of cyberbullying and one category of non-cyberbullying. The proposed framework is designed to enhance cyberbullying detection performance with low computational complexity for short text sequences. This work focuses on a lightweight attention-enhanced CNN that is trained from scratch, without the need for pretrained transformer models or larger architectures. The model is designed to be an efficient solution for detecting cyberbullying in short texts, where the CNN feature extraction is combined with residual attention refinement.

The key contributions of this work are outlined below:

- (1) A lightweight base deep learning framework for accurately classifying cyberbullying into five categories;
- (2) Optimized hyperparameter settings for cyberbullying detection in short tweet sequences;
- (3) An enhanced attention-embedding mechanism that highlights salient contextual patterns of bullying-related short tweet sequences;
- (4) Facilitates deployment on resource-constrained devices while maintaining a low inference latency for detecting cyberbullying.

2 Related Works

Recently, extensive studies have explored the applications of deep learning in text analysis, including the detection of cyberbullying. Specifically, CNNs, RNNs, and their derivatives have demonstrated a high capability for extracting contextual and semantic information from short and ambiguous social media posts, prompting extensive research to improve automated cyberbullying detection [14–16].

2.1 Deep Learning Approaches for Text Classification

Among existing deep learning algorithms, CNN-based models have exhibited high performance in extracting local textual patterns and decoding discriminative terms from short text posts [15,17]. Based on these principles, Minaee et al. [18] presented a study involving improving the performance of convolutional networks through the strategic inclusion of a dynamic pooling process, which focuses on extracting fine-grained features from text paragraphs. In a related context, some studies have focused on increasing the depth of convolutional networks to improve textual feature extraction, but this comes with increased computational complexity and a higher risk of overfitting [19]. Therefore, many studies have focused on building a baseline model capable of scaling up by strategically adding enhancement modules to further improve performance while maintaining low computational complexity [20–22]. Meanwhile, some studies have employed other equivalent networks, such as recurrent neural networks (RNNs) [23,24], long short-term memory (LSTMs) [22,25], gated recurrent units (GRUs) [26,27], and bidirectional long short-term memory (Bi-LSTMs) [27,28], which are widely used in natural language processing due to their capability to handle sequential relationships and contextual information [29]. In practice, numerous studies have demonstrated the efficiency of these equivalent networks, even in detecting cyberbullying. However, these recurrent structures require more computing resources and present challenges in processing limited-word posts [14]. Based on these findings and challenges, some studies have focused on incorporating various types of attention mechanisms. This strategy contributes to enhancing the effectiveness of text segment analysis and selectively focusing on targeted keywords and patterns [30]. Accordingly, Wang et al. [31] investigated the inclusion of multi-scale attention mechanisms in deep learning models to focus on underlying textual features. Attention mechanisms can be precisely engineered to focus on specific linguistic features in emotional expressions and abusive language patterns by controlling the weight of each word within sentences [32].

2.2 Deep Learning for Cyberbullying Detection

Advances in deep learning-based text classification have positively impacted the development of highly specialized cyberbullying detection systems. To validate this claim, researchers Atoum [33] conducted a comparative study evaluating the performance of a deep CNN network and machine learning-based networks such as the Support Vector Machine (SVM) and the Naïve Bayes algorithm for cyberbullying detection. The deep learning model demonstrated superiority over traditional machine learning methods. In a related context, the authors Badjatiya et al. [34] conducted another comparative study involving CNN, LSTM, and FastText models for detecting hate speech embedded in texts. The study concluded that the CNN model with random initialization outperformed the other models. More deeply, the authors Iwendi et al. [35] experimented with a wide range of deep networks, specifically RNNs, LSTMs, GRUs, and Bi-LSTMs. In addition, they proposed a modified LSTM model with improved gate structures, resulting in increased classification accuracy for cyberbullying, despite its significant increase in computational complexity. To further investigate this aspect, Wu et al. [17] conducted a series of experiments to explore the ability of models such as CNNs, recurrent CNNs (RCNNs), CNN-LSTMs, and FastText to efficiently extract textual relationships and achieve rich contextual representations through hybridization between models. To this extent, it can be concluded that recurrent and hybrid models have improved the classification performance of textual content, whereas many current studies focus on models based on memory-dependent structures as a strategy to capture sequential dependencies, which in turn increases computational complexity. Additionally, text-based social media posts frequently include a restricted set of highly discriminative terms and phrases that are not suitable for intricate sequence modelling. On the other hand, previous studies have demonstrated the effectiveness of attention mechanisms embedded in deep networks, as they enhance the extraction of prominent textual features associated with cyberbullying behavior. Despite this, attentional strategies have not been sufficiently explored.

2.3 Recent Trends in Cyberbullying Detection

Recently, there has been an increasing number of studies dedicated to the application of attention-based, multi-modal, and reasoning-oriented deep learning approaches for the detection of cyberbullying. Mali et al. used a stacked Bi-GRU attention model along with BERT for the detection of cyberbullying behavior in social media texts [36]. For the same purpose, Ismail et al. presented an adaptive and hierarchical fusion-based multilingual multimodal framework for detecting cyberbullying behavior in social media text-image [37]. Other recent research has expanded the scope of cyberbullying research to include reasoning-based detection, contextual interpretation, and mitigation [38,39]. These works emphasize the importance of conveying semantics, intent, emotion, sarcasm, ambiguity, and context in the conversation. While there are robust advances in these studies, many recent studies are based on complex recurrent, transformer-based, multi-modal or reasoning-based architectures. Compared to that, the current study employs a light attentive CNN-based approach for the efficient detection of numerous classes of cyberbullying at low computational cost by using text-based information.

Therefore, there is a growing need for computationally efficient frameworks capable of successfully highlighting prominent indicators of cyberbullying while maintaining high classification accuracy in short social media posts.

3 Methodology

This section details the general methodology proposed in this work, including the dataset used, text preprocessing techniques, base CNN model architectures, hyperparameter optimization, and attention mechanism embedding. Fig. 1 illustrates the workflow for the proposed cyberbullying detection framework.

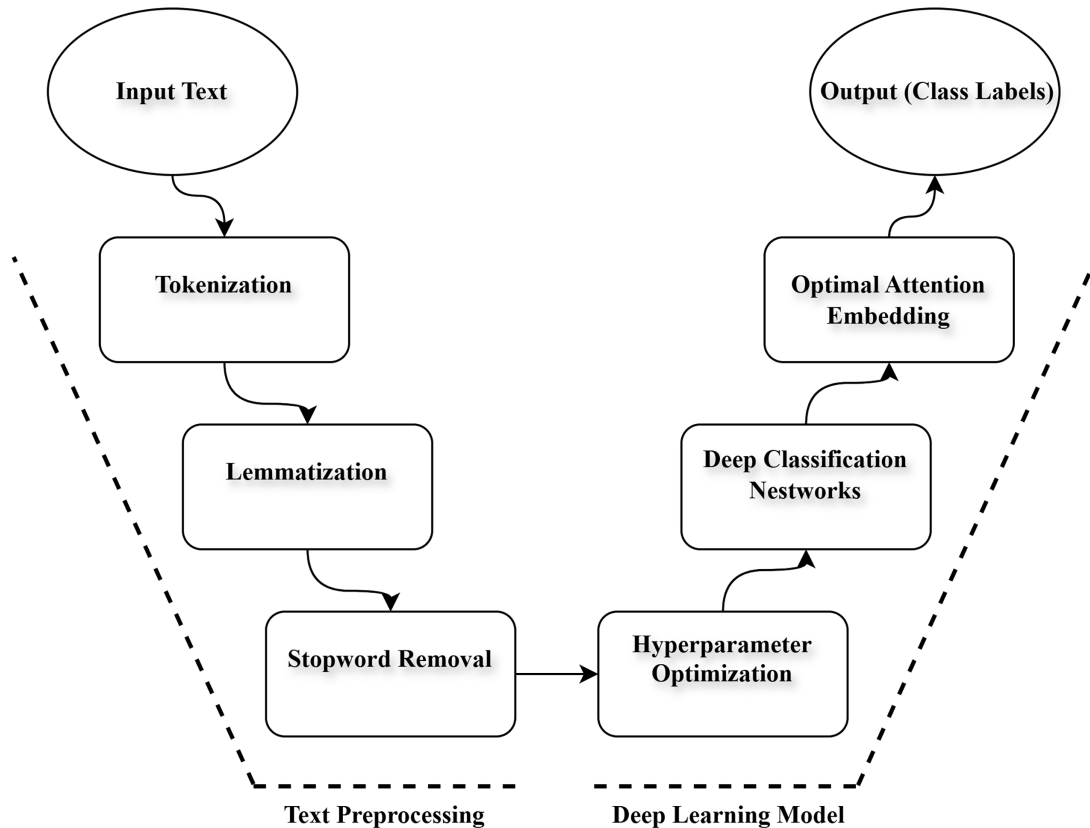


Figure 1: General workflow of the proposed attention-based cyberbullying detection framework.

Initially, a public dataset containing cyberbullying tweets is downloaded and organized into a CSV file. This data was then processed and split into a 75% training set and a 25% independent test set. The validation set was formed by allocating 20% of the training set during hyperparameter tuning, and the independent test set was not used in this process but was only used at the end of the process for a performance estimate. This procedure was used to minimize the chance of leakage of validation data and provide overly optimistic performance predictions. After that, these tokenized sets are used as input for deep neural networks. Based on the results obtained, the optimal baseline architecture is determined by evaluating several configuration scenarios involving varying depth and different CNN components. Subsequently, the hyperparameters are systematically optimized using a greedy search to consider all possible outcomes, where each hyperparameter is assessed in turn, and the value that performs the best response is preserved before the next parameter is optimized. Finally, an attention mechanism is embedded into the model to selectively focus on the most relevant words, thereby enhancing feature representation and improving cyberbullying classification accuracy.

3.1 Cyberbullying Tweets Dataset

The original dataset employed in this paper comprises 47,692 tweets annotated into six classes: age, gender, religion, ethnicity, other cyberbullying, and non-cyberbullying [40]. The dataset is publicly available and can be accessed through Kaggle at <https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification>. This database is characterized by a nearly balanced number of samples per category. Based on this, the original distribution was maintained, eliminating the need to create an additional oversampling or perform random selection. This balanced class distribution enables better training and more reliable performance evaluation. Representative examples of tweets from each category are provided in Table 1 to illustrate the linguistic characteristics and diversity of cyberbullying expressions present in the dataset.

Table 1: Sample cyberbullying tweets classified according to their respective categories.

Category	Sample of Tweets	# of Samples
Not cyberbullying	I can't stand split keyboards. doesn't work well with MMOs.	7945
Gender	It's obvious the blonde bimbos make better TV according to Channel 7. Shit cooks.	7973
Religion	Don't forget, she always tells you that she is a Christian Woman. Yeah RIGHT!!	7998
Age	Stop thats my nickname at school that people always use to bully me :(((	7992
Ethnicity	Fuck you doris ya dumb nigger.	7961
Other cyberbullying	AGeekCalledBob ick. DV can happen to anyone.	7823

Personal information of Twitter users has been removed to preserve privacy, and the tweets cannot be traced back to the original users. The ground-truth annotations were provided by an independent third party, whereby conflicts in category assignment were resolved through consensus to ensure a fair annotation

process. Based on this, tweets that do not fall into any of the predefined categories are labelled as “other Cyberbullying,” while tweets that do not contain any bullying intent are labelled as “non-cyberbullying.”

3.2 Text Preprocessing

The raw tweets are almost noisy and unstructured data that need pre-processing before model training. The data preprocessing in this paper is designed to convert raw text into a clean and standardized format that is suitable for training deep neural network models. Specifically, the text was fully converted to lowercase, and all punctuation was removed [41]. After that, the resulting text was tokenized into individual words, and stopword removal was applied to remove the frequently used terms that are not carrying any contextual information and can increase the training time [42,43]. The context-sensitive words, however, were kept, especially the negation words “no”, “not”, and “never”, because they might alter the meaning of short abusive or non-abusive messages. Subsequently, the lemmatization process based on the NLTK library is applied to convert the processed words into their original forms while keeping semantic significance [44].

3.3 Base Convolutional Neural Network Architecture

The first step in developing a foundation for subsequent improvements is to design a baseline model. In this paper, four fundamental network architectures were constructed, which are designated as A, B, C, and D, and then evaluated systematically. The simplest model, Architecture (A), consists of an input layer and an embedding layer, followed by a shallow convolutional layer and a pooling layer. The extracted features are then passed through a flattening layer, a dropout layer, and a fully connected layer, before producing the final output layer. The full network design of the Architecture (A) is depicted in Fig. 2.

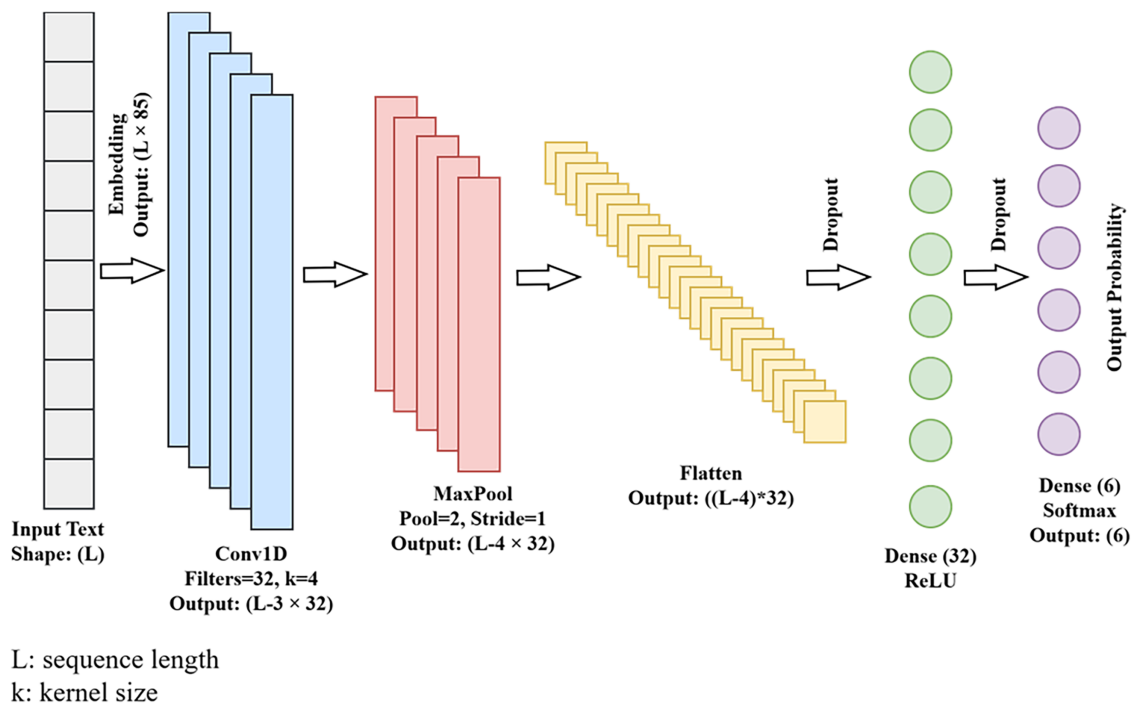


Figure 2: Architecture of baseline 1D CNN network used for cyberbullying text classification.

Architectures (B, C, and D) have the same structure as architecture (A) but are deeper. Architecture (B) includes two convolutional-pooling layers, while C and D include three and four layers, respectively.

These four architectures are evaluated using a standardized configuration. The hyperparameters were fixed to ensure a fair comparison; in addition, the input text was also standardized at 280 characters, in accordance with the maximum tweet length.

3.4 Hyperparameter Tuning

An effective hyperparameter optimization method will improve the learning rate of the model, whilst avoiding the possibility of local extremum [45]. In this work, the hyperparameter tuning was achieved using a greedy search strategy. Based on this strategy, each hyperparameter was experimented with individually, and the best optimised value obtained at each stage was passed on to the subsequent optimisation stage. This sequential tuning approach was used to minimise computational requirements while ensuring a systematic and repeatable optimisation process. The order of the tuning parameters was chosen to start by stabilizing the training process with the optimizer and learning rate, then to tune the capacity of the model representation by tuning the size of the convolutional filters and the nodes of the fully connected layers. Finally, the number of epochs and dropout rate were optimised to enhance regularisation and training time. The experiment covered a set of hyperparameters including learning rate, filter size, optimiser, dropout rate, epochs, and the number of nodes within the fully connected layer [46]. Table 2 lists the summary of the hyperparameter search space considered in this study. In each step of the greedy search, the value of the hyperparameter that maximizes the validation accuracy was chosen, and the validation loss was served used as a secondary criterion.

Table 2: Hyperparameter search space for the baseline CNN architecture.

Hyperparameter	Values Tested
Optimizer	Adam, Adamax, Adadelta, Adagrad, Nadam, SGD
Learning rate	Default, 0.01, 0.001, 0.0001, 0.00001
Convolutional filter size	8, 16, 32, 64, 128
Dropout rate	0.1, 0.2, 0.3, 0.4, 0.5
Fully connected layer nodes	10, 20, 32, 64, 128
Epoch	5, 10, 25, 35, 70, 80, 90, 100

Typically, each experimental run began with the following procedure: The preprocessed tweets were passed through the 1D CNN model, which was composed of an embedding layer that mapped each token to an 85-dimensional vector representation. The embedding layer input dimension was calculated based on the size of the vocabulary, while the output dimension was kept constant in all experiments. To illustrate the hyperparameter tuning process, the filter size was chosen as an example, with filter sizes of 8, 16, 32, 64, and 128, as shown in Fig. 3.

3.5 Attention Mechanism for Cyberbullying Detection

The embedding of an attention mechanism into the deep learning model is performed so that focused features can be better extracted from relevant segments of the input sequences. More specifically, these modules provided higher weight or importance to some related cyberbullying-related patterns or phrases. The importance of incorporating these mechanisms increases, especially when dealing with short social media text inputs containing a very limited number of ambiguous words that might contain bullying. To illustrate how different weights are assigned to each term in Twitter texts when attention mechanisms are incorporated, Fig. 4 provides some illustrative examples adapted from a previous study [32].

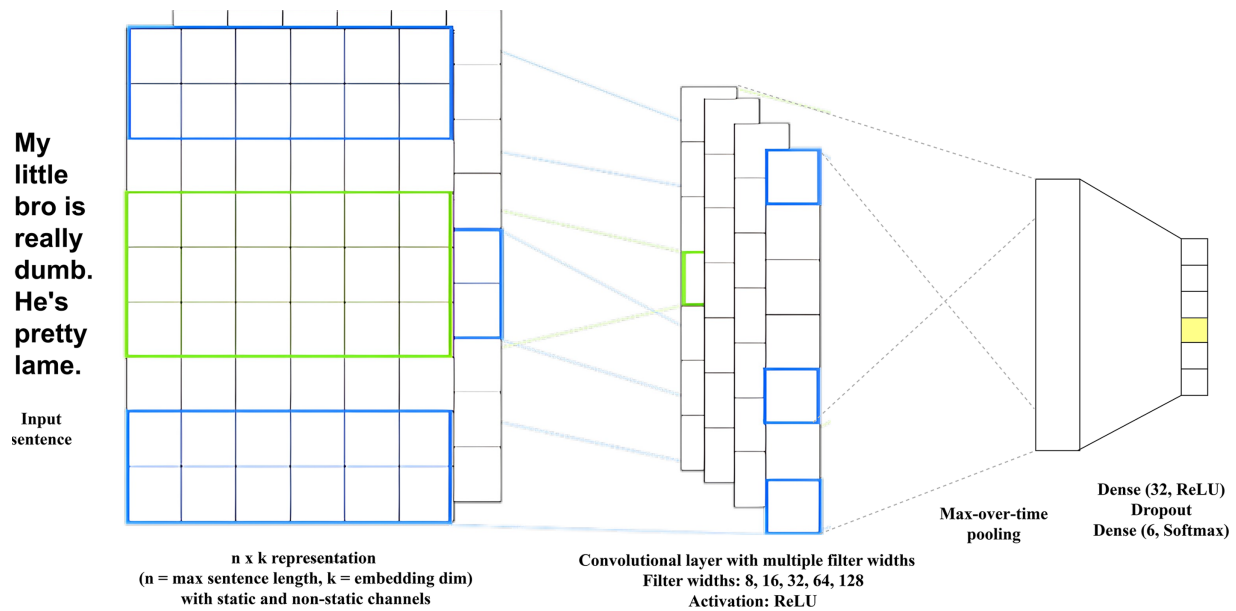


Figure 3: Tokenized tweets representation and feature flow through the 1D CNN model.

RT @user: y'all be having routines for everything except that ugly a## attitude
 That ni##er food in the cafe today was disgusting
 You must air his show live in showroom. Guests can hear Rush say 's#ut, FemiNazi, nags...
 Ummm do a currency conversion i#iot
 Stupid lil b##ch done pissed me off. I ain't let nobody make me this mad in months
 @user Bi##h :) lol
 You're a bit of a do#k, aren't you?
 My litto bro is really d#mb. He's pretty l#me.
 RT @user: All religions are not the same. Islam is more violent than other religions.
 Why US? Ask you king first to ban India you idiot.
 GAG! Another fake bullsh##ing spewing Clinton... I'm feeling nauseous!!
 those blond girls make me crazy, they are so annoying

Figure 4: Illustrative examples of attention weights assigned to input tweets using the attention mechanism, adapted from [32].

The figure is provided as a qualitative illustration to show how attention mechanisms can help identify important textual cues, but is not intended as quantitative interpretability evidence for the proposed model. Thus, the interpretability claim of this work is restricted to demonstrating the general attention weighting behavior and not a complete token-level explanation of model decisions. Dark shading is associated with higher weights and terms clearly linked to bullying, while lighter colors are used for ordinary terms. This visualisation can explain how the deep learning model enhances the classification process by focusing on semantic signals in ambiguous texts.

Multiple network configurations were systematically assessed to identify the most optimal integration of the attention mechanism. Some of the configurations that were explored include variations in the depth of the encoder and decoder, activation functions, convolutional filter sizes, and training epochs. The proposed attention branch is composed of three encoder–decoder stages. The sequence resolution is reduced in each encoder stage through MaxPooling1D with a pool size of 2 and a stride of 2, and each decoder stage is done by UpSampling1D with a size of 2 to restore the sequence resolution. The Conv1D operations are performed with the stride argument set to 1 and with the same padding, and retain the sequence dimension at each resolution level. The complete architecture of the proposed attention mechanism is depicted in Fig. 5.

The mathematical operations underlying the proposed attention mechanism illustrated in Fig. 5 are formulated as follows. Given a tokenised tweet $w = [\omega_1, \dots, \omega_L]$, the embedding layer transforms the input into an 85-dimensional feature representation:

$$E = [e(\omega_1), e(\omega_2), \dots, e(\omega_L)] \in \mathbb{R}^{B \times L \times 85} \quad (1)$$

where $e(\omega_i)$ represents the embedding vector of the i -th token. The attention module separately generates trunk features and a non-negative feature-attention mask:

$$T = f_{\text{trunk}}(E), M = \text{ReLU}(f_{\text{mask}}(E)) \quad (2)$$

here, the corresponding functions are the convolutional trunk branch, the encoder–decoder attention branch, and the final convolutional refinement operation, respectively. The trunk branch retains the original sequence resolution and generates a feature map $T \in \mathbb{R}^{B \times L \times 32}$. Simultaneously, the length of the sequence is being shrunk by three successive MaxPooling1D operations with pool size = 2, stride = 2, which leads to resolutions of $L/2, L/4, L/8$. The bottleneck is then at $L/8$.

The decoder then applies three UpSampling1D (size = 2) operations to recover the sequence dimension at the respective resolution levels, and the Conv1D (stride = 1, same padding) operations keep the sequence length the same for each resolution level. Thus, the final attention feature map M is of size $B \times L \times 32$, which is the same size as the trunk feature map T . The trunk features and feature-attention mask are combined through element-wise multiplication and residual addition, followed by convolutional refinement:

$$F_{\text{att}} = f_{\text{refine}}[T + (T \odot M)] = f_{\text{refine}}[T \odot (1 + M)] \quad (3)$$

where $\odot$ denotes element-wise multiplication. This residual formulation preserves the trunk representation while adaptively amplifying salient cyberbullying-related features. This design was used because short tweets tend to contain a few very discriminative words or phrases to cyberbully. The encoder–decoder attention branch further enriches the local CNN features and highlights related textual information, and the residual connection is used to maintain the original representation and avoid filtering out useful information. This design aligns with the lightweight design goals of the proposed framework, where convolutional encoder–decoder operations are used in the attention branch. Finally, the attention-refined features are concatenated with the features extracted by the parallel CNN branch:

$$F_{\text{fused}} = \text{Concat}[f_{\text{CNN}}(E), F_{\text{att}}] \quad (4)$$

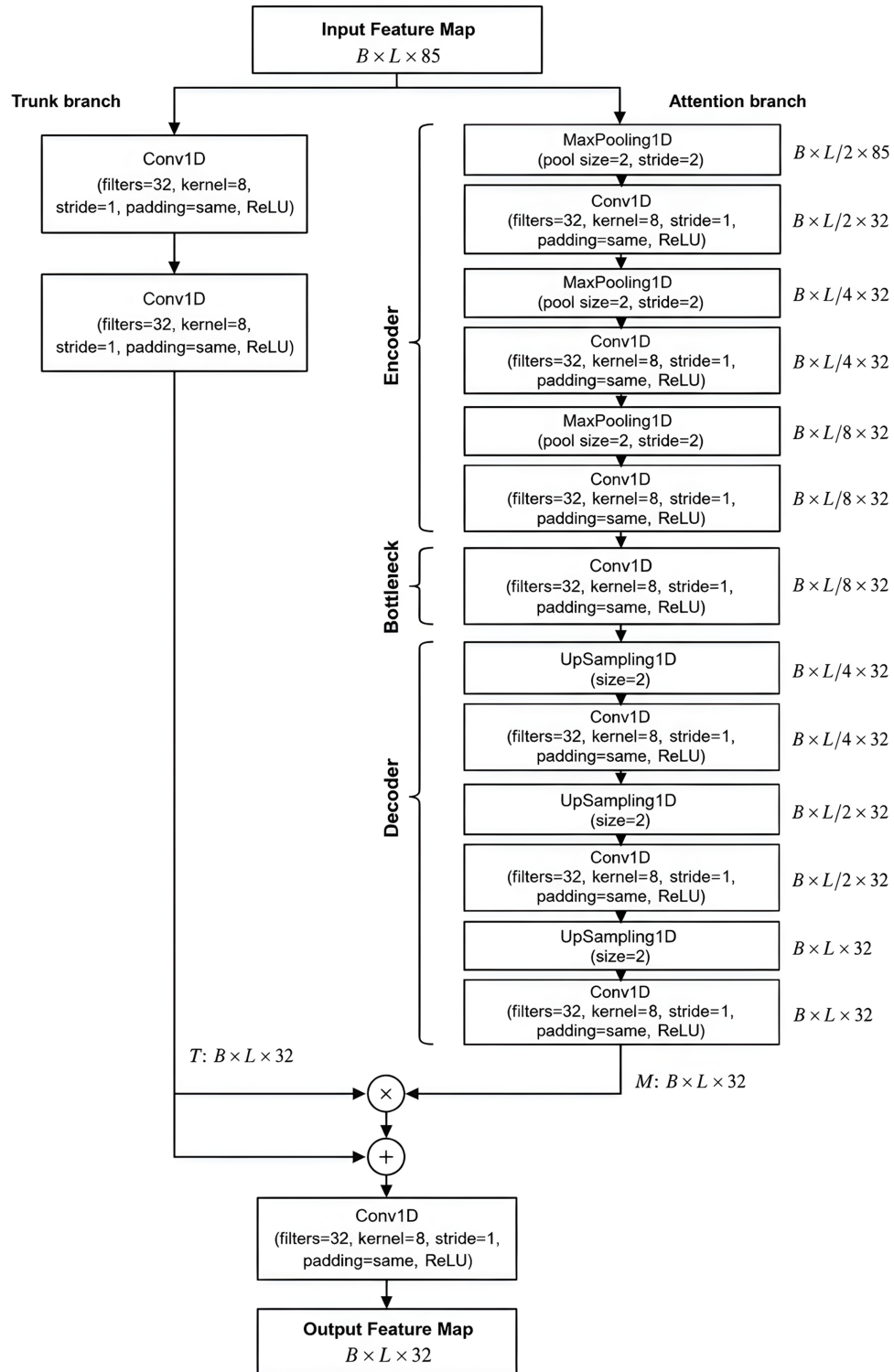


Figure 5: Architecture of the proposed attention mechanism embedded within a CNN framework for highlighting cyberbullying-related textual features. The batch size is denoted as B ; the padded sequence length is denoted as L ; the number of features in the trunk is denoted as T ; the number of features in the attention is denoted as M .

The resulting concatenated representation is then fed to the classification pipeline, including convolution, pooling, flattening, dropout, and fully connected layers to predict the final class. From a mathematical perspective, the listed equations demonstrate the precise mechanism of the proposed model and how it enhances distinctive patterns related to cyberbullying while preserving the general textual information used in the context of the texts. According to this formulation, an additional evaluation procedure was systematically performed to identify the most effective attention configuration. The MaxPooling1D and UpSampling1D settings are fixed architectural components of the attention branch and were therefore not included in the hyperparameter search space. The evaluation uses a defined search space strategy to examine a set of architectural configurations and training hyperparameters, as listed in [Table 3](#). Eventually, the optimal configuration was determined by the highest accuracy score and the lowest loss.

Table 3: Hyperparameter search space for the attention mechanism.

Hyperparameter	Values Tested
Encoder/decoder layers	2, 3, 4
Convolutional filter size	16, 32, 64, 128
Activation function	Sigmoid, ReLU, tanh
Batch normalization	With, Without
Epoch	5, 15, 35, 55

4 Experiments and Results

This section reports the experimental results and discusses the findings related to a set of experiments. The experiments included building the baseline classification model, optimizing its hyperparameters, embedding attentional modules, and benchmarking the enhanced model with the existing architectures. All these experiments were conducted using the 3.8.5 Python environment and TensorFlow 2.10 to ensure fair comparison and consistency.

4.1 Baseline CNN Architecture

The first set of experiments focused on the evaluation of the performance of proposed architectures using the same hyperparameter settings. Five independent runs were performed for each baseline architecture with different random seeds, and the results are presented as mean $\pm$ standard deviation. The architecture (A), which consists of an embedding layer, a single convolutional layer, a pooling layer, a flattening layer, a dropout layer, a fully connected layer, and an output layer, achieved the highest performance in comparison to the other baseline architectures, as shown in [Table 4](#).

Table 4: Evaluation and comparison of baseline architectures for cyberbullying detection performance over five independent runs.

Architecture	Loss	Accuracy
A—CNN layer (One)	1.6066 $\pm$ 0.0039	0.7944 $\pm$ 0.0035
B—CNN Layers (Two)	1.7780 $\pm$ 0.0050	0.7835 $\pm$ 0.0040
C—CNN layers (Three)	2.1811 $\pm$ 0.0044	0.7865 $\pm$ 0.0038
D—CNN layers (Four)	2.4784 $\pm$ 0.0035	0.7750 $\pm$ 0.0040

Based on the listed results, it can be concluded that increasing the depth of convolutional networks negatively impacts overall network performance. Accordingly, the shallower design (architecture A) is considered as a baseline for the subsequent experiments as it is highly suited for analyzing short texts of 280 characters.

4.2 Baseline Hyperparameter Tuning

After the baseline model was adopted, the next step was to conduct an extensive series of systematic experiments using a greedy search strategy to obtain the best hyperparameters. In more detail, the experiments included six testable hyperparameters, where the most probable values were experimented with for each hyperparameter individually to reach the highest model performance. The optimal value for each hyperparameter resulting from the experiments was then fixed and used in the subsequent experiments as a constant. The applied search space and the optimum values obtained for each hyperparameter are summarised in [Table 5](#).

Table 5: Hyperparameter optimisation results for the baseline CNN architecture over five independent runs.

Hyperparameter	Search Range	Optimal Value	Accuracy	Loss
Optimizer	Adam, Adamax, Adadelta, Adagrad, Nadam, SGD	SGD	0.8006 ± 0.0033	0.4724 ± 0.00502
Learning rate	Default, 0.01, 0.001, 0.0001, 0.00001	Default	0.8006 ± 0.0040	0.4724 ± 0.0033
Convolutional filter size	8, 16, 32, 64, 128	32	0.8142 ± 0.0051	0.4687 ± 0.0049
Dropout rate	0.1, 0.2, 0.3, 0.4, 0.5	0.5	0.8210 ± 0.0043	0.4965 ± 0.0042
Fully connected layer nodes	10, 20, 32, 64, 128	32	0.8226 ± 0.0040	0.4356 ± 0.0035
Epochs	5, 10, 25, 35, 70, 80, 90, 100	35	0.8226 ± 0.0039	0.4356 ± 0.0044

4.3 Performance Evaluation of the Optimised Baseline Architecture

Based on the optimal combination obtained in the previous section, the enhanced baseline model achieved a classification accuracy of 0.8226, a loss score of 0.4356, a precision of 0.8250, a recall of 0.8150, and an F1-score of 0.8200. The optimised CNN yielded an accuracy of 0.8226 and a loss of 0.4356 compared to the baseline CNN (Architecture A) with an accuracy of 0.7944 and a loss of 1.6066, as reported in [Table 4](#). [Fig. 6](#) shows the training and validation loss curves for the enhanced model.

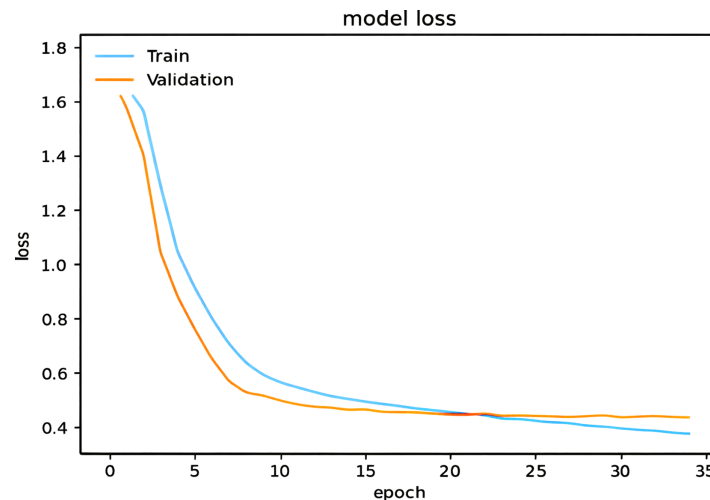


Figure 6: Loss convergence curves of the enhanced baseline model with optimal hyperparameter configurations.

4.4 Attention Mechanism Embedding

Following up on previous investigations, an additional set of experiments was conducted to improve the efficiency of the attention mechanisms embedded in the baseline model. These experiments involved five tunable parameters, with all potential values for each hyperparameter assessed progressively using the greedy search approach, as summarised in [Table 6](#).

Table 6: Hyperparameter optimisation results for the attention mechanism.

Hyperparameter	Search Range	Optimal Value	Accuracy	Loss
Encoder-decoder layers	2, 3, 4	3	0.8077	0.7574
Activation function	Sigmoid, ReLU, tanh	ReLU	0.8104	0.5924
Batch normalization	With, Without	Without	0.8151	0.5751
Convolutional filter size	16, 32, 64, 128	32	0.8270	0.5305
Epoch	5, 15, 35, 55	15	0.8311	0.4880

Based on the listed results, the optimised attention-based CNN model achieved an accuracy of 0.8311 and a loss score of 0.4880. Statistically, this model outperformed the optimised CNN without an attention module, which achieved an accuracy of 0.8226 and an F1-score of 0.8200. Furthermore, a direct paired-design statistical comparison was performed on the same five experimental runs presented in [Table 7](#) to check whether these improvements were statistically greater than the random variation of repeated training. To ensure that any variation due to data splitting and training stochasticity was eliminated, the same data partition and random seed were used for each paired run with an optimised CNN and an optimised attention + CNN.

Table 7: An ablation study of the attention mechanism impact over five paired experimental runs.

Model	Accuracy (Mean $\pm$ SD)	Δ Accuracy	p -Value	F1-Score (Mean $\pm$ SD)	Δ F1	p -Value
Optimized CNN without Attention	0.8226 $\pm$ 0.0025	–	–	0.8200 $\pm$ 0.0029	–	–
Optimized Attention + CNN	0.8311 $\pm$ 0.0031	+0.0085	<0.001	0.8217 $\pm$ 0.0030	+0.0017	0.0074

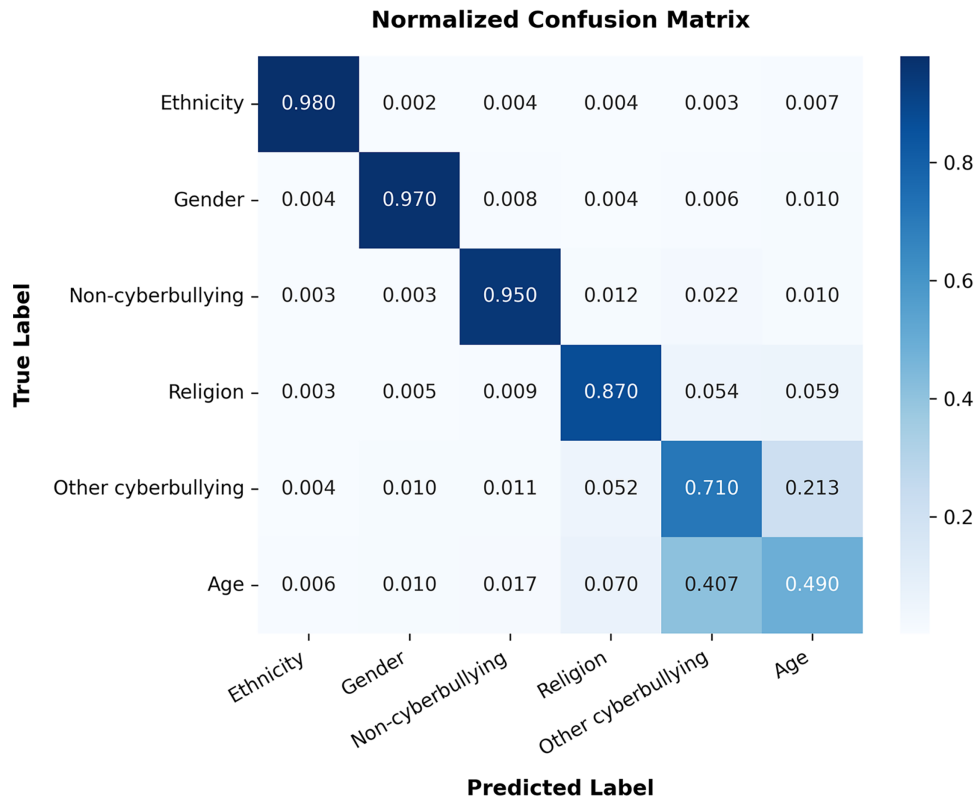
The ablation analysis shows that by including the attention mechanism, the mean classification accuracy is improved from 0.8226 to 0.8311, which is an absolute gain of 0.0085. This improvement was confirmed to be statistically significant by the paired t -test ($p < 0.001$). The mean F1-score improved from 0.8200 to 0.8217, and had a statistically significant difference between paired runs ($p = 0.0074$). The findings suggest an improvement in the attention module that is consistent across run-to-run training variability. However, the improvement noted for the F1 score is comparatively small, and it should thus be considered to be a statistically significant but limited improvement. In a more detailed analysis, the improved model achieved an overall precision, recall rate, and F1 score of 0.83, as shown in [Table 8](#).

These results indicate improved performance for the proposed system after incorporating the attention mechanisms. However, lower-class-wise performance was observed for “Age” and “other cyberbullying”, which could be attributed to the semantic overlap between these two categories.

Table 8: Class-wise performance of the optimized attention-based CNN model.

Class	F1-Score	Recall	Precision
Ethnicity	0.98	0.98	0.98
Gender	0.97	0.97	0.97
Non-cyberbullying	0.95	0.95	0.95
Religion	0.87	0.87	0.86
Other cyberbullying	0.65	0.71	0.59
Age	0.55	0.49	0.62
Overall	0.83	0.83	0.83

To further analyse the differences in the performance of the optimised attention-based CNN model in each category, the normalised confusion matrix is used, as shown in Fig. 7.

**Figure 7:** Normalized confusion matrix of the optimized attention-based CNN model.

The values of the diagonal elements are high, with 0.980 for Ethnicity, 0.970 for Gender, 0.950 for Non-cyberbullying, and 0.870 for Religion. Other cyberbullying and Age were found to have lower performances (0.710 and 0.490, respectively). The primary disagreement was between these two categories, with 40.7% of Age samples misclassified as Other cyberbullying, and 21.3% of Other cyberbullying samples misclassified as Age. This is because there is a cross-over of abusive terms in these categories, and some of these terms are specific to the context. Hence, the lower F1-scores of Age and Other cyberbullying can be explained by the fact that they share semantic aspects and have divergent linguistic patterns.

4.5 Performance Benchmarking with State-of-the-Art Models

The inter-class confusion implies a need for more effective class separation for categories with similar meanings. Hierarchical classification, contrastive representation learning, or class-specific data augmentation may be useful to enhance discrimination between Age and Other Cyberbullying by separating general abusive content from target-specific abusive expressions.

To provide additional evidence of the effectiveness of the optimized baseline architecture (without the attention), it was compared against various commonly used deep learning models such as RNN, LSTM, GRU, and BiLSTM as reported in Table 9. Based on the analysis of the obtained results, it can be easily concluded that the proposed network, without the addition of attention mechanisms, clearly outperformed the other standard architectures. This superior performance is followed by that of the referenced CNN model, which had the highest baseline accuracy of 0.7944, followed closely by the BiLSTM model with an accuracy of 0.7938, and the GRU model had the lowest accuracy of 0.7133 among the evaluated models.

Table 9: Performance benchmarking of baseline deep learning models.

Model	Loss	Accuracy	Precision	Recall	F1-Score
CNN [47]	0.4724	0.7944	0.7800	0.7867	0.7833
RNN [48]	0.5483	0.7630	0.4960	0.4980	0.4970
LSTM [11]	0.5373	0.7691	0.8167	0.7967	0.8066
GRU [12]	0.8205	0.7133	0.8017	0.8033	0.8025
BiLSTM [49]	0.5041	0.7938	0.7367	0.7500	0.7433
Optimized-CNN (Without Attention)	0.4356	0.8226	0.8250	0.8150	0.8200

These findings demonstrate the effectiveness of CNN architectures in learning discriminative features from short tweets with relatively reduced computational complexity. On the other hand, the performance of the collection of classification models was evaluated after adding the attention mechanisms, as summarized in Table 10. For these investigations, experiments were run five times with varying random number seeds, and the average and standard deviation of each metric were obtained. This iterative testing gives a more accurate measurement of model stability than if a single run is reported. These procedures ensure the robustness and reliability of the evaluation process. For multi-class evaluation, Precision, Recall, and F1-score were calculated independently for each class and then macro-averaged over the six classes, so the macro F1-score reported is not directly calculated from the macro Precision and macro Recall scores. To make the statistical comparison, paired *t*-tests were carried out on the five experimental runs used to obtain the F1-scores for the Optimized Attention + CNN model as the baseline.

Table 10: Performance benchmarking of attention-embedded deep learning models (mean $\pm$ standard deviation).

Model	Loss	Accuracy	Precision	Recall	F1-Score	<i>p</i> -Value
Attention Mechanism + CNN	0.5927 $\pm$ 0.0048	0.8010 $\pm$ 0.0028	0.8100 $\pm$ 0.0022	0.8117 $\pm$ 0.0030	0.8100 $\pm$ 0.0037	$p < 0.001$
Attention Mechanism + LSTM	0.7526 $\pm$ 0.0033	0.7898 $\pm$ 0.0042	0.8083 $\pm$ 0.0035	0.8067 $\pm$ 0.0034	0.8067 $\pm$ 0.0033	$p < 0.001$
Attention Mechanism + GRU	0.7822 $\pm$ 0.0038	0.7860 $\pm$ 0.0039	0.8200 $\pm$ 0.0041	0.8133 $\pm$ 0.0033	0.8117 $\pm$ 0.0041	$p < 0.001$
Attention Mechanism + BiLSTM	0.5145 $\pm$ 0.0035	0.7801 $\pm$ 0.0034	0.8033 $\pm$ 0.0039	0.8067 $\pm$ 0.0030	0.8067 $\pm$ 0.0050	$p < 0.001$
Attention Mechanism + Bi GRU	0.5398 $\pm$ 0.0040	0.7929 $\pm$ 0.0048	0.8033 $\pm$ 0.0052	0.8050 $\pm$ 0.0040	0.7983 $\pm$ 0.0048	$p < 0.001$
Optimized Attention + CNN	0.4880 $\pm$ 0.0043	0.8311 $\pm$ 0.0050	0.8350 $\pm$ 0.0035	0.8250 $\pm$ 0.0049	0.8217 $\pm$ 0.0036	–

The listed results revealed the superiority of the final proposed model over the other models, achieving an accuracy of 0.8311 and an F1 score of 0.8217. These findings reflect the ability of the proposed model to efficiently classify different types of cyberbullying. The largest performance gain was observed for the optimized attention-based CNN architecture, where the accuracy increased by an average of 4.62%. Although recurrent architectures such as LSTM, GRU, and BiLSTM are effective in capturing sequential dependencies, the results indicate that short tweets often contain a limited number of highly discriminative words and phrases that can be effectively modeled using convolutional operations combined with attention mechanisms.

To evaluate the statistical reliability, the paired t -test metric was calculated using the obtained results from the five-run F1-scores, considering the Optimized Attention + CNN model as the reference. F1 score is chosen as a metric since it is a balance between precision and recall in a multi-class classification, which indicates that the proposed model's improvements. All comparisons yielded a p -value less than 0.001 were statistically significant at the 0.05 level. All comparisons yielded $p < 0.001$, indicating statistically significant differences between the proposed model and the compared models.

For the broader benchmarking analysis, paired t -tests were additionally conducted using the five-run F1-scores of the Optimized Attention + CNN as the reference against the other attention-enhanced architectures reported in Table 10. These comparisons are distinct from the direct attention ablation analysis reported in Table 7, which specifically evaluates the effect of adding the proposed attention mechanism to the Optimized CNN. All benchmarking comparisons yielded $p < 0.001$, indicating statistically significant differences between the proposed model and the evaluated attention-enhanced architectures. The notation $p < 0.001$ simply means the reporting threshold for each comparison; it does not mean that the same value of p was found for all comparisons. In addition, the paired comparisons showed statistically significant differences; however, the results are viewed with caution as the number of paired experimental runs is limited (only five), which reduces the statistical power and generalizability of the significance test and thus complements the significance test results with the size of the differences in the F1-scores.

The investigation process was further extended to check the discriminative power of the proposed model by using the class-wise receiver operating characteristic (ROC) analysis as shown in Fig. 8. The mean AUC for the model was 0.965 across the 6 classes, with AUC values being 1.00 for Gender, 1.00 for Ethnicity, 0.98 for Religion, 0.90 for Age, 0.92 for other cyberbullying, and 0.99 for non-cyberbullying. The results show that most of the cyberbullying classes discriminated well, with comparatively lower performance for the Age and Other cyberbullying classes.

While the proposed framework has promising results on English tweets, it cannot be directly applied to the multilingual detection of cyberbullying. To extend the model to multilingual contexts, preprocessing needs to be language-aware, language models need to be multilingual, or it needs to be based on a cross-lingual transfer learning approach to understand socially mixed language and abusive expressions across languages. Additionally, the proposed model may be affected by the domain shift phenomenon and the changing nature of the language used in social media. The model might suffer from poor generalisation abilities when it is deployed to new online platforms with emerging slang, platform-specific terms, and hostile modifications of text. Thus, there is a need for periodic updating of the model, robust representations of texts, and augmentation-based strategies for enhancing resistance to new patterns of cyberbullying in practical deployment.

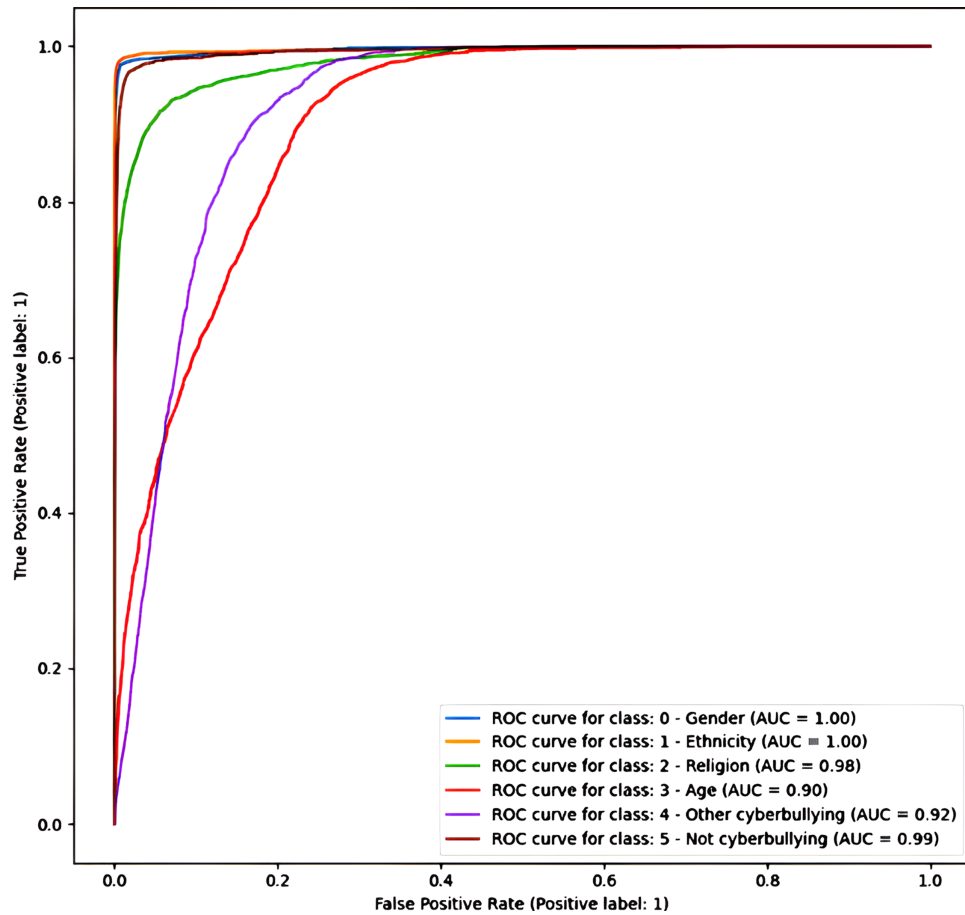


Figure 8: Class-wise ROC curves of the optimised attention-based CNN model for the six cyberbullying categories.

4.6 Computational Complexity Analysis

In addition to classification performance, computational efficiency is essential for the practical deployment of cyberbullying detection systems, particularly in real-time social media monitoring. Therefore, the evaluated attention-based models were compared in terms of model parameters, floating-point operations (FLOPs) per sample, and inference time. All measurements were obtained using the same test data and an NVIDIA Tesla P100-PCIE-16GB GPU. Inference latency was measured using a batch size of one after 50 warm-up iterations and is reported as the mean $\pm$ standard deviation over five runs. To ensure consistency across models, FLOPs were computed using TensorFlow graph-based profiling with a batch size of one in the forward-pass computational graph.

As shown in Table 11, the proposed Attention-based CNN model required only 1.520 million parameters and 107.461 million FLOPs per sample, representing reductions of 30.8% and 71.7%, respectively, compared with the conventional Attention + CNN. It maintained a comparable inference time of 2.5349 ± 0.0416 ms per sample with the highest throughput of 394.49 sample/S, which is considered substantially faster than the recurrent variants, which required 7.3–10.7 ms. Furthermore, the proposed model required a training time of 0.2683 min/epoch with a peak GPU memory usage of 0.218 GB on the Tesla accelerator. These results demonstrate a favorable balance between classification performance and computational efficiency, supporting real-time cyberbullying detection.

Table 11: Computational complexity comparison of the evaluated attention-based models using an NVIDIA GPU.

Model	Parameters (M)	FLOPs (M/Sample)	Inference Time (ms/Sample)	Throughput (Sample/S)
Attention + CNN	2.198	380.283	2.6303 $\pm$ 0.0329	380.18 $\pm$ 4.76
Attention + LSTM	2.187	329.363	7.2983 $\pm$ 0.0923	137.02 $\pm$ 1.73
Attention + GRU	2.175	329.363	7.2852 $\pm$ 0.2532	137.26 $\pm$ 4.77
Attention + BiLSTM	1.985	328.860	10.7392 $\pm$ 0.0420	93.12 $\pm$ 0.36
Attention + BiGRU	1.961	328.860	10.7134 $\pm$ 0.0400	93.34 $\pm$ 0.35
Proposed attention-based CNN	1.520	107.461	2.5349 $\pm$ 0.0416	394.49 $\pm$ 6.47

In terms of deployment, the lightweight design of the proposed model is more suitable for resource-constrained systems, including mobile devices, edge computing systems, and real-time social media monitoring systems. Reduced parameter counts and FLOPs can reduce memory usage and computational cost, and the low inference latency allows for fast detection of harmful content. In online moderation systems, the proposed model can be used as an early filtering module that flags potentially harmful textual content. This initial alarm can be processed more effectively by the subsequent contextual analysis or by human experts. However, despite its lightweight design and practical deployment potential, ethical issues must be considered when implementing an automated system to detect cyberbullying. Issues such as false positives, demographic/linguistic bias, user privacy, transparency, and human oversight are still significant concerns. The proposed framework, therefore, should be interpreted as a decision support system and not as a standalone moderation system, especially in cases of sensitive online settings where the wrong decision can impact users' expression and platform fairness.

5 Conclusion

This work introduced an attention-enhanced one-dimensional CNN framework for cyberbullying detection in short social media posts. By combining a compact CNN backbone with nested attention mechanisms, the proposed framework achieved high scores in cyberbullying detection while keeping low computational complexity. Different CNN network configurations were extensively analyzed and evaluated, reporting that the shallowest CNN architecture with a single convolutional layer achieved the best performance. This simplified structure, which does not rely on the computationally heavier recurrent or transformer-based architectures, is compatible with the limited length of tweet-based inputs. Accordingly, a greedy hyperparameter optimization was performed to adopt the baseline model with the optimum performance. The optimized baseline model achieved an accuracy of 0.8226 and a loss of 0.4356, outperforming the non-optimized model by 3.55%. Additional performance gains were achieved through the inclusion of the attention mechanisms. These modules enabled the model to focus selectively on ambiguous patterns relevant to cyberbullying. Building upon this improvement, the optimized attention-based CNN model achieved the highest classification accuracy of 0.8311, outperforming the optimized baseline model by 1.03%. Meanwhile, the proposed model was benchmarked against five referenced deep learning models guided by attention mechanisms, demonstrating the efficacy of these added mechanisms across all models. Beyond classification accuracy, the proposed model was also compact, with reductions of 30.8% in parameters and 71.7% in FLOPs, supporting its lightweight design and its capability to be used for low-latency cyberbullying monitoring. It is important to note that the reported inference latency was obtained on an NVIDIA Tesla P100 GPU and should not be interpreted as an indicator of deployment performance when using mobile, edge, or CPU-based resource-constrained platforms. Despite the improved performance, several limitations

remain for this work. The most significant is its focus solely on text-based posts, neglecting other post types such as images and emojis, which are prevalent on most social media platforms. Furthermore, it relies on a single spatial attention mechanism without exploring dual and stacked attention mechanisms. The greedy search was efficient and repeatable, but might not be able to capture interactions between hyperparameters. Random search, Bayesian optimization, and cross-validation-based tuning are to be explored in the future to make the model more robust. Future studies can be expanded to assess pretrained transformer-based models, including BERT, RoBERTa, DeBERTa, and DistilBERT. Such models need an experimental design with external pretraining, transformer-specific tokenization, and special fine-tuning. This will enable a wider range of analysis of the accuracy–efficiency trade-off in cyberbullying detection.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by Universiti Kebangsaan Malaysia under Dana Padanan Kolaborasi-Geran Pembiayaan Sepadan with grant number DPK-GPS-JORDAN-2024-021 and ISIF-Asia Grant with grant number KK-2022-018 (M-202205-01161).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Mohammad J. M. Zedan and Edwin Rangka Ardhana; methodology, Edwin Rangka Ardhana and Mohd Asyraf Zulkifley; software, Mohammad J. M. Zedan; validation, Edwin Rangka Ardhana and Mohammad J. M. Zedan; formal analysis, Mohammad J. M. Zedan; investigation, Mohammad J. M. Zedan; resources, Mohd Asyraf Zulkifley; data curation, Siti Raihanah Abdani; writing original draft preparation, Mohammad J. M. Zedan; writing review and editing, Edwin Rangka Ardhana, Siti Raihanah Abdani and Mohd Asyraf Zulkifley; visualization, Mohammad J. M. Zedan; supervision, Mohd Asyraf Zulkifley; project administration, Mohd Asyraf Zulkifley; funding acquisition, Mohd Asyraf Zulkifley. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset can be downloaded from <https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kee DMH, Anwar A, Vranjes I. Cyberbullying victimization and suicide ideation: the mediating role of psychological distress among Malaysian youth. *Comput Hum Behav.* 2024;150(5):108000. doi:10.1016/j.chb.2023.108000.
2. MacAulay PJR, Betts LR, Stiller J, Kellezi B. Bystander responses to cyberbullying: the role of perceived severity, publicity, anonymity, type of cyberbullying, and victim response. *Comput Hum Behav.* 2022;131(5):107238. doi:10.1016/j.chb.2022.107238.
3. Li J, Wu Y, Hesketh T. Internet use and cyberbullying: impacts on psychosocial and psychosomatic wellbeing among Chinese adolescents. *Comput Hum Behav.* 2023;138(1):107461. doi:10.1016/j.chb.2022.107461.
4. Nikolaou D. Does cyberbullying impact youth suicidal behaviors? *J Health Econ.* 2017;56(5):30–46. doi:10.1016/j.jhealeco.2017.09.009.
5. Nisha M, Jebathangam J. Deep KNN based text classification for cyberbullying tweet detection. In: *Proceedings of the 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART); 2022 Dec 16–17; Moradabad, India.* New York, NY, USA: IEEE; 2023. p. 1550–4. doi:10.1109/SMART55829.2022.10047054.
6. Vitman O, Kostiuk Y, Sidorov G, Gelbukh A. Sarcasm detection framework using context, emotion and sentiment features. *Expert Syst Appl.* 2023;234(4):121068. doi:10.1016/j.eswa.2023.121068.

7. Zedan MJM, Saleh A, Harun HH, Zulkifley MA. Machine learning methods in cyberbully detection: a concise review. In: Proceedings of the 2024 10th International Engineering Conference on Advances in Computer and Civil Engineering (IEC); 2024 Oct 23–24; Erbil, Iraq. New York, NY, USA: IEEE; 2025. p. 84–9. doi:10.1109/IEC61018.2024.11064257.
8. Ali Mohammed Alhasani D, Patel KA. A survey on text classification using deep learning approaches. In: Illés Z, Verma C, Gonçalves PJS, Singh Y, editors. Proceedings of International Conference on Recent Innovations in Computing. Singapore: Springer Nature; 2025. p. 377–94. doi:10.1007/978-981-96-6034-6_25.
9. Mohammed AJ, Al-Hameed W. Cyberbullying detection using deep learning: RoBERTa and CNN approaches. In: Proceedings of the Third International Conference on Emerging Trends in AI and Computational Technologies (ICONEST 2025); 2025 Nov 18–19; Tuban, Indonesia. p. 303–11. doi:10.1117/12.3107968.
10. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. Eng Appl Artif Intell. 2022;115:105151. doi:10.1016/j.engappai.2022.105151.
11. Ghosh S, Chaki A, Kudeshia A. Cyberbully detection using 1D-CNN and LSTM. In: Sabut SK, Ray AK, Pati B, Acharya UR, editors. Proceedings of International Conference on Communication, Circuits, and Systems. Singapore: Springer; 2021. p. 295–301. doi:10.1007/978-981-33-4866-0_37.
12. Luo Y, Zhang X, Hua J, Shen W. Multi-featured cyberbullying detection based on deep learning. In: Proceedings of the 2021 16th International Conference on Computer Science & Education (ICCSE); 2021 Aug 17–21; Lancaster, UK. p. 746–51. doi:10.1109/iccse51940.2021.9569270.
13. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62. doi:10.1016/j.neucom.2021.03.091.
14. Balakrisnan V, Kaity M. Cyberbullying detection and machine learning: a systematic literature review. Artif Intell Rev. 2023;56(1):1375–416. doi:10.1007/s10462-023-10553-w.
15. Elizar E, Zulkifley MA, Muharar R, Zaman MHM, Mustaza SM. A review on multiscale-deep-learning applications. Sensors. 2022;22(19):7384. doi:10.3390/s22197384.
16. Fitro A, Fitro A, Wibowo MA, Widodo CE. Automatic detection of cyberbullying on text, image, and video: a systematic literature review. J Sisfokom Sistem Informasi Dan Komputer. 2025;15(1):75–80. doi:10.32736/sisfokom.v15i01.2542.
17. Wu H, Liu Y, Wang J. Review of text classification methods on deep learning. Comput Mater Continua. 2020;63(3):1309–21. doi:10.32604/cmc.2020.010172.
18. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning—based text classification: a comprehensive review. ACM Comput Surv. 2022;54(3):1–40. doi:10.1145/3439726.
19. Sadiq S, Mehmood A, Ullah S, Ahmad M, Choi GS, On BW. Aggression detection through deep neural model on twitter. Future Gener Comput Syst. 2021;114(2):120–9. doi:10.1016/j.future.2020.07.050.
20. Iqbal T, Qureshi S. The survey: text generation models in deep learning. J King Saud Univ Comput Inf Sci. 2022;34(6):2515–28. doi:10.1016/j.jksuci.2020.04.001.
21. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: a survey. Mech Syst Signal Process. 2021;151:107398. doi:10.1016/j.ymssp.2020.107398.
22. Wang J, Peng B, Zhang X. Using a stacked residual LSTM model for sentiment intensity prediction. Neurocomputing. 2018;322(2):93–101. doi:10.1016/j.neucom.2018.09.049.
23. Schaefer AM, Udluft S, Zimmermann HG. Learning long-term dependencies with recurrent neural networks. Neurocomputing. 2008;71(13–15):2481–8. doi:10.1016/j.neucom.2007.12.036.
24. Majumdar A, Gupta M. Recurrent transform learning. Neural Netw. 2019;118(9):271–9. doi:10.1016/j.neunet.2019.07.003.
25. Qin C, Chen L, Cai Z, Liu M, Jin L. Long short-term memory with activation on gradient. Neural Netw. 2023;164(2):135–45. doi:10.1016/j.neunet.2023.04.026.
26. Kaur S, Singh S, Kaushal S. Deep learning-based approaches for abusive content detection and classification for multi-class online user-generated data. Int J Cogn Comput Eng. 2024;5:104–22. doi:10.1016/j.ijcce.2024.02.002.

27. Raj C, Agarwal A, Bharathy G, Narayan B, Prasad M. Cyberbullying detection: hybrid models based on machine learning and natural language processing techniques. *Electronics*. 2021;10(22):2810. doi:10.3390/electronics10222810.
28. Li W, Qi F, Tang M, Yu Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*. 2020;387(4):63–77. doi:10.1016/j.neucom.2020.01.006.
29. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys D Nonlinear Phenom*. 2020;404(8):132306. doi:10.1016/j.physd.2019.132306.
30. Wu P, Li X, Ling C, Ding S, Shen S. Sentiment classification using attention mechanism and bidirectional long short-term memory network. *Appl Soft Comput*. 2021;112(2):107792. doi:10.1016/j.asoc.2021.107792.
31. Wang R, Li Z, Cao J, Chen T, Wang L. Convolutional recurrent neural networks for text classification. In: *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*; 2019 Jul 14–19; Budapest, Hungary. p. 1–6. doi:10.1109/ijcnn.2019.8852406.
32. Fang Y, Yang S, Zhao B, Huang C. Cyberbullying detection in social networks using Bi-GRU with self-attention mechanism. *Information*. 2021;12(4):171. doi:10.3390/info12040171.
33. Atoum JO. Cyberbullying detection neural networks using sentiment analysis. In: *Proceedings of the 2021 International Conference on Computational Science and Computational Intelligence (CSCI)*; 2021 Dec 15–17; Las Vegas, NV, USA. New York, NY, USA: IEEE; 2022. p. 158–64. doi:10.1109/CSCI54926.2021.00098.
34. Badjatiya P, Gupta S, Gupta M, Varma V. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th International Conference on World Wide Web Companion—WWW' 17 Companion*; 2017 Apr 3–7; Perth, Australia. p. 759–60. doi:10.1145/3041021.3054223.
35. Iwendi C, Srivastava G, Khan S, Maddikunta PKR. Cyberbullying detection solutions based on deep learning architectures. *Multimed Syst*. 2023;29(3):1839–52. doi:10.1007/s00530-020-00701-5.
36. Mali MK, Pawar RR, Shinde SA, Kale SD, Mulik SV, Jagtap AA, et al. Automatic detection of cyberbullying behaviour on social media using stacked bi-gru attention with BERT model. *Expert Syst Appl*. 2025;262(21):125641. doi:10.1016/j.eswa.2024.125641.
37. Ismail WS, Ullah H, Adnan M, Ullah F. Multilingual multimodal cyberbullying detection through adaptive and hierarchical fusion. *Array*. 2026;29(3):100689. doi:10.1016/j.array.2026.100689.
38. Nadeem M, Chen H. RaSA-BoDX: a meta-cognitive reasoning framework for cyberbullying language detection and mitigation using multi-agent systems. *Knowl Based Syst*. 2026;339(4):115639. doi:10.1016/j.knosys.2026.115639.
39. Nadeem M, Chen H. Breaking the classification bubble: rethinking cyberbullying detection and detoxification as knowledge-guided reasoning systems. *Comput Sci Rev*. 2026;62(4):101009. doi:10.1016/j.cosrev.2026.101009.
40. Wang J, Fu K, Lu CT. SOSNet: a graph convolutional network approach to fine-grained cyberbullying detection. In: *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*; 2020 Dec 10–13; Atlanta, GA, USA. p. 1699–708. doi:10.1109/bigdata50022.2020.9378065.
41. Uysal AK, Gunal S. The impact of preprocessing on text classification. *Inf Process Manag*. 2014;50(1):104–12. doi:10.1016/j.ipm.2013.08.006.
42. Philipo AG, Sarwatt DS, Ding J, Daneshmand M, Ning H. Cyberbullying detection: exploring datasets, technologies, and approaches on social media platforms. *ACM Comput Surv*. 2026;58(7):1–35. doi:10.1145/3785654.
43. Kathuria A, Gupta A, Singla RK. A review of tools and techniques for preprocessing of textual data. In: *Proceedings of the Computational Methods and Data Engineering*. Singapore: Springer; 2020. p. 407–22. doi:10.1007/978-981-15-6876-3_31.
44. Sarica S, Luo J. Stopwords in technical language processing. *PLoS One*. 2021;16(8):e0254937. doi:10.1371/journal.pone.0254937.
45. Raiaan MAK, Sakib S, Fahad NM, Al Mamun A, Rahman MA, Shatabda S, et al. A systematic review of hyperparameter optimization techniques in convolutional neural networks. *Decis Anal J*. 2024;11(1):100470. doi:10.1016/j.dajour.2024.100470.
46. Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing*. 2020;415(1):295–316. doi:10.1016/j.neucom.2020.07.061.

47. Al-Ajlan MA, Ykhlef M. Deep learning algorithm for cyberbullying detection. *IJACSA*. 2018;9(9):199–205. doi:10.14569/ijacsa.2018.090927.
48. Markoski F, Zdravevski E, Ljubešić N, Gievska S. Evaluation of recurrent neural network architectures for abusive language detection in cyberbullying contexts. In: *Proceedings of the 17th International Conference on Informatics and Information Technologies (CIIT 2020)*; 2020 May 8–9; Online.
49. Cunha Sergio G, Lee M. Stacked DeBERT: all attention in incomplete data for text classification. *Neural Netw*. 2021;136(3):87–96. doi:10.1016/j.neunet.2020.12.018.