SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Detection of Cyberbullying on Facebook Twitter (X) Using Bi-Directional Long Short-Term Memory and Extreme Gradient Boost Algorithms

Joseph Adebayo Ojeniyi , Yusuf Adamu Mohammed, Abdulkadir Onivehu Isah , Peter Chizaramuekpere Anyaora * , Fasola Olusanjo , Andrew Uduimoh and Meshach Baba

Cyber Security Science Department, Federal University of Technology Minna, Minna 920101, Nigeria

* Correspondence: p.anyaora@futminna.edu.ng

Received: 26 October 2025; **Revised:** 5 December 2025; **Accepted:** 12 March 2026; **Published:** 1 July 2026

Abstract: The social networking sites have transformed digital communication but have simultaneously enabled the escalation of harmful online behaviors, particularly cyberbullying. This recurring form of digital aggression can lead to serious emotional and psychological harm, including anxiety, depression, and in severe cases, self-inflicted injury or suicidal behavior. The timely identification and prevention of cyberbullying have become an essential focus of current research. Although numerous machine learning techniques have been applied to detect abusive content, many continue to face challenges such as inefficient kernel tuning, extended training durations, and reduced predictive accuracy. To address these limitations, this study presents a hybrid deep learning architecture that integrates a Bidirectional Long Short-Term Memory (BiLSTM) network with the Extreme Gradient Boosting (XGBoost) algorithm to improve contextual awareness and classification accuracy. The proposed framework was trained and evaluated on datasets collected from Facebook and X (formerly Twitter), capturing diverse linguistic and behavioral characteristics of user interactions. Experimental results indicate that the BiLSTM–XGBoost hybrid model outperforms conventional classifiers by effectively managing context representation, adaptive learning, and class imbalance. The model achieved 97% accuracy, 95% precision, 92% recall, and an F1-score of 96%, confirming its robustness and efficiency for cyberbullying detection in dynamic social media environments. The study helps educational institutions, online platforms and legal frameworks provide insights into how to better identify cyberbullying in real-world scenarios. The study's high recall ensures that cyberbullies are easily identified and it enhances the understanding of how combining multiple models can lead to better performance in cyberbullying detection.

Keywords: Cyberbullying; Deep Learning; BiLSTM; XGBoost; Machine Learning; Social Media

1. Introduction

Over the past decade, cyberbullying has evolved into a serious social and technological issue, particularly affecting young individuals who actively engage with online platforms such as Facebook and X (formerly Twitter) [1] (p. 1). It encompasses deliberate acts of intimidation, humiliation or harassment carried out through digital communication channels, frequently leading to psychological distress and emotional instability among victims [2] (p. 1). Such behaviors often manifest as the dissemination of false information, unauthorized sharing of personal materials, use of derogatory or hateful expressions, impersonation, or repetitive online aggression intended to inflict

emotional harm. The anonymity and widespread accessibility of digital platforms enable offenders to target individuals effortlessly, which can escalate into long-term psychological trauma and in extreme circumstances, suicidal behavior [3] (pp. 1–7).

As the prevalence of cyberbullying intensifies, researchers have increasingly focused on developing computational strategies capable of detecting and mitigating harmful online interactions. Machine learning (ML) techniques have emerged as an effective approach for analyzing extensive volumes of social media data to uncover underlying linguistic and behavioral patterns [4] (p. 1). However, building a highly reliable detection system remains challenging since individual algorithms often struggle to capture the nuanced contextual and semantic diversity inherent in user-generated online content [5].

To overcome these challenges, ensemble-based learning methods have become increasingly valuable, offering a mechanism to merge the predictive capabilities of multiple algorithms to improve accuracy and stability. Specifically, stacked learning frameworks integrate the outcomes of various classifiers into a consolidated model, producing a more consistent and dependable prediction process [6] (p. 2). By combining different algorithms, such ensemble approaches mitigate the weaknesses of single models and enhance the overall reliability of cyberbullying detection systems [7]. Beyond only causing grief, cyberbullying causes severe psychological injury. It causes emotional disengagement, long-term fear, and low self-esteem. Online attacks frequently cause victims to obsess on them, which impairs their ability to think clearly, perform well in school, and control their emotions. Shame, social distrust, and an elevated risk of self-harm and suicidal thoughts are all consequences of public humiliation. Long-term research reveals persistent susceptibilities to symptoms of trauma, anxiety, and depression [7]. Additionally, research indicates that long-term stress from cyberbullying may impact brain development, making it harder to regulate emotions and make decisions. The body of research indicates that cyberbullying has a significant negative impact on identity, feelings, and long-term mental health [3].

In this research, a stacking-inspired hybrid architecture is introduced for detecting cyberbullying on Facebook and X (Twitter) platforms. The proposed model was evaluated alongside state-of-the-art architectures, including BERT (Bidirectional Encoder Representations from Transformers), a leading Natural Language Processing (NLP) model renowned for its bidirectional text encoding, which interprets word meaning by analyzing both preceding and succeeding contextual information [8] (p. 2). While BERT and other transformer-based systems have shown strong performance, they still face limitations such as increased sensitivity to hyperparameter configurations and higher false-positive rates in certain contexts [9] (p. 2).

To address these issues, the present study introduces an enhanced hybrid model that integrates Bidirectional Long Short-Term Memory (BiLSTM) with the Extreme Gradient Boosting (XGBoost) algorithm. This combination strengthens contextual learning, improves feature extraction, and increases the system's adaptability to linguistic diversity. The structure of this paper is organized as follows: Section Two outlines a literature review of existing studies, Section Three outlines the materials and methodological framework, Section Four outlines, presents and analyzes the experimental findings, and Section Five outlines the conclusions of the study with key insights and recommendations for further research. This study was able to:

I. Design an ensemble model architecture for detection of cyberbullying.
II. Then, the implementation of the ensemble model design architecture.

2. Literature Review

2.1. Cyberbullying Detection

Cyberbullying refers to the intentional use of electronic communication platforms such as social media, instant messaging applications, and online forums to harass, intimidate, or emotionally harm others [1] (p. 2). Such online hostility can manifest through abusive messages, rumor dissemination, or derogatory and humiliating language directed at individuals or groups. Victims of these behaviors often endure prolonged emotional distress, which may lead to anxiety, depression, social withdrawal, or, in extreme circumstances, self-harm and suicidal ideation. Consequently, governments, academic institutions, and social media organizations have introduced technical safeguards and policy interventions to reduce exposure to online abuse and foster safer digital environments.

Recognizing the urgency of this social problem, researchers have increasingly explored machine learning (ML) as a means to automate the detection of online aggression. ML-based approaches can process vast quantities of

user-generated data to identify linguistic and behavioral cues indicative of bullying or harassment [10]. According to Dadvar et al. [11] (p. 2), this challenge can be formulated as a supervised classification task, where the objective is to discriminate between bullying and non-bullying content while minimizing misclassification rates. Using Recurrent Neural Networks (RNNs), their system achieved an accuracy of 82.2%, successfully distinguishing harmful messages from benign ones.

Traditional ML and early Natural Language Processing (NLP) techniques, however, often fall short in capturing the subtleties of human expression particularly sarcasm, coded language, and cultural nuances [12] (p. 2). To overcome these shortcomings, recent research has turned toward Large Language Models (LLMs) including advanced systems such as ChatGPT, Claude, Gemini, and Mistral which employ prompt-based learning to achieve superior interpretive and contextual understanding. These modern LLMs have shown notable improvements in both semantic comprehension and predictive reliability across datasets from social media environments like Facebook and Reddit, with Claude 3 and Mistral exhibiting the highest consistency [12] (p. 3).

Building upon these advancements, Agbaje and Afolabi [13] (p. 3) designed a multimodal hybrid framework that integrates Convolutional Neural Networks (CNNs) for processing visual data with Recurrent Neural Networks (RNNs) for textual information, complemented by sentiment analysis. This multimodal fusion of linguistic and visual features improved the detection of aggressive behavior compared to text-only approaches. Similarly, Alqahtani and Ilyas [14] (p. 3) introduced an ensemble learning model that combined K-Nearest Neighbors (KNN), Decision Trees, Random Forest, Linear Support Vector Classifier (SVC), and Logistic Regression, together with feature representation techniques such as Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe. Their configuration achieved a 94% accuracy rate, identifying TF-IDF as the most efficient feature extraction method.

The proliferation of abusive and aggressive interactions continues to grow across major online platforms, including Facebook, Instagram, and X (formerly Twitter) [14] (p. 3). The anonymity and informal communication style prevalent on these platforms characterized by abbreviations, sarcasm, and slang pose additional challenges for automated detection systems. As a result, researchers have developed ensemble-based models that leverage the complementary strengths of multiple algorithms to improve contextual understanding and predictive precision.

A decade-long review conducted identified recurring challenges in the field, including dataset imbalance [15] (pp. 3, 4), domain-specific bias, and limited generalizability across languages. While many algorithms perform well on English-language corpora, their effectiveness diminishes in multilingual or low-resource languages such as Arabic and Bengali. Moreover, the scarcity of labeled datasets continues to constrain progress in this area. The same review confirmed that Support Vector Machines (SVMs) and Logistic Regression (LR) remain dependable baseline models when combined with TF-IDF or BoW features, often reaching 94–97% accuracy. However, modern deep learning architectures such as CNN-LSTM, Gated Recurrent Unit (GRU), and Bidirectional LSTM (BiLSTM) networks consistently outperform these traditional algorithms, with accuracy values approaching 99.9% under controlled testing conditions.

To address linguistic diversity, Brakas and Alanezi [16] (p. 3) have proposed a study targeting Arabic-language social media interactions among university students in Mosul, Iraq. Using manually annotated Facebook comments obtained through the Graph API and APIFY, their research revealed a significant presence of cyber aggression within Arabic-speaking communities. After text normalization and tokenization, features were extracted via TF-IDF, and the data were divided into 80–20 train-test partitions. The Logistic Regression classifier achieved high predictive accuracy, confirming its adaptability for underrepresented languages.

Further developments have also explored multi-class and multilingual classification strategies. For instance, Reynolds et al. [17] (pp. 3–6) implemented a multi-label neural network to categorize Facebook posts into five distinct abuse categories non-bullying, sexual harassment, threats, trolling, and religious discrimination achieving 87.91% accuracy. Similarly, Jadhav et al. [18] (p. 4) employed a Gated Recurrent Unit (GRU) model to analyze Bengali-language posts across thematic categories such as politics, hate speech and religion, attaining 70.1% accuracy.

Earlier lexicon-based methods have also contributed to this field. For example, Atoum [19] (p. 3) designed a three-tier dictionary-driven system for offensive language detection. Then, Bharadwaj et al. [20] (p. 3) proposed a gender-aware SVM framework for MySpace datasets, which demonstrated improved recall, precision, and F-measure over traditional classifiers. Furthermore, Vijayakumar et al. [21] (p. 3) presented a multilingual CNN in-

corporating T5-Sentence embeddings, yielding high F1-scores across several languages. Similarly, BiLSTM models trained with FastText embeddings exhibited improved accuracy when applied to informal or non-standard linguistic contexts, such as Bengali social media communication.

2.2. Cyberbullying across Online Platforms

Social media platforms have become vital components of modern digital interaction and numerous empirical studies have examined the patterns and consequences of cyberbullying across multiple online environments including Twitter, YouTube, Ask.fm, and various chat-based systems. For instance, Twitter facilitates rapid, opinion-driven communication through short posts limited to 280 characters, which provides a valuable resource for analyzing abusive behavior. Using a dataset of tweets containing offensive or harmful content, sentiment-based classification and categorization into four distinct groups was applied [22], negative with bullying intent, negative without bullying intent, positive, and neutral. Their proposed model achieved an accuracy rate of 67.3%.

A complementary study explored by Gupta et al. [23] (p. 4), a semi-anonymous question and answer site, to assess how user anonymity affects online aggression. The researchers examined over 30,000 user profiles using both network analysis metrics and content-based features, uncovering that individual with limited social interaction or engagement were more susceptible to harassment. Similarly, Reynolds et al. [17] (pp. 3, 4) analyzed real-world chat transcripts from online predator investigations and successfully identified linguistic markers such as grooming, isolation, and trust manipulation, achieving 93% accuracy in detecting predatory behavior.

Further work by Shi et al. [24] (p. 4) distinguished between chat-oriented and discussion-oriented digital environments. A comparative study involving Kongregate (a chat-driven platform) and MySpace (a thread-based platform) revealed that TF–IDF weighting produced superior classification performance over N-gram methods, achieving an F-score of 0.481 and precision of 0.394.

In addition to supervised methods, researchers have also examined unsupervised learning approaches for cyberbullying detection. For example, Akhter et al. [25] (p. 4) proposed a method that integrates textual indicators and social interaction features to detect abusive communication without labeled data. While this approach yielded varying levels of performance across datasets, it demonstrated potential for broader application in real-world contexts. Likewise, Reynolds et al. [17] (pp. 3, 4) experimented with hybrid feature representations combining TF–IDF and sentiment-based N-grams, confirming improved accuracy when textual and semantic features were combined.

Expanding this research direction, Al-Hashedi et al. [26] (p. 4) presented a machine learning and natural language processing (ML–NLP) hybrid framework that compared Bag-of-Words (BoW) and TF–IDF features across four learning algorithms to detect abusive language. Also, Dewani et al. [9] (p. 4) introduced a binary classification model trained on user tweets and profile meta-data to detect emotional states such as depression or aggression. A meta-analysis of sixteen text-based studies later demonstrated that both feature extraction techniques and dataset scale substantially influence model accuracy and generalization [26] (p. 4).

Another major contribution came from Ali and Syed [27] (p. 4), who trained multiple ML models on a global Twitter dataset using both TF–IDF and Word2Vec embeddings to enhance contextual learning. A subsequent study analyzed Facebook comment threads, categorizing content into racism, harassment, and shaming through a Multinomial Naïve Bayes (MNB) classifier. Comparative evaluations revealed that multi-class classification yielded richer insights into online aggression than binary models.

Furthermore, Hande et al. [28] (p. 4) proposed a cross-linguistic detection framework for multilingual datasets, then, Islam et al. [29] (p. 4) explored hybrid sentiment-recognition systems combining English and Arabic corpora. These multilingual initiatives addressed growing concerns regarding the linguistic diversity of cyberbullying. However, challenges such as irony and sarcasm interpretation remain persistent. To mitigate these, Dewani et al. [9] (p. 4) combined sentiment and profanity detection modules, achieving 74% accuracy and an F1-score of 0.74.

Arabic-language research has also advanced significantly. For instance, Kumar and Sachdeva [30] (p. 4) and Muneer et al. [31] (p. 4) contributed to the SemEval-2019 shared task on offensive language detection, while Raj et al. [32] (pp. 3, 4) developed an integrated message classification and network monitoring system to track harassment trends. In related work, Hande et al. [28] (p. 4) proposed a filtering architecture designed to identify and suppress offensive posts on social media platforms.

The importance of cross-lingual and multicultural research was emphasized by Keni et al. [33] (p. 4), who observed that earlier studies largely focused on English-based datasets. To address this gap, Hani et al. [34] (p. 4)

designed a Support Vector Machine (SVM)-based "Concise" system optimized for detecting aggression in Instagram posts, while Thorat et al. [35] (p. 4) introduced "Samurai", an intelligent detection engine leveraging automation to classify harmful content in real time. Advances in neural architectures, such as the CNN–BiLSTM configuration, have further enhanced computational efficiency through optimized ReLU and Sigmoid activation functions [3] (pp. 2–4).

Finally, Mehendale et al. [36] (p. 4) implemented a Random Forest model trained on Reddit and Kaggle datasets, achieving an Area Under the Curve (AUC) score of 0.90 and a precision rate of 0.89 despite inherent data imbalance. A broader taxonomy of online aggression including trolling, stalking, impersonation, doxing, and exclusion was summarized by Raj et al. [32] (p. 4), as the major categories of digital harassment prevalent across social media ecosystems.

3. Materials and Methods

3.1. Datasets and Input Layer

According to Zhong et al. [37] (p. 5), the architecture developed in this research adopts a Bidirectional Long Short-Term Memory (BiLSTM) neural framework, carefully configured to identify instances of cyberbullying within diverse social media environments. To assess the robustness and generalizability of the model, two separate datasets were employed during the evaluation phase. **Figure 1** presents the complete methodology integrating BiLSTM and XGBoost for cyberbullying detection.

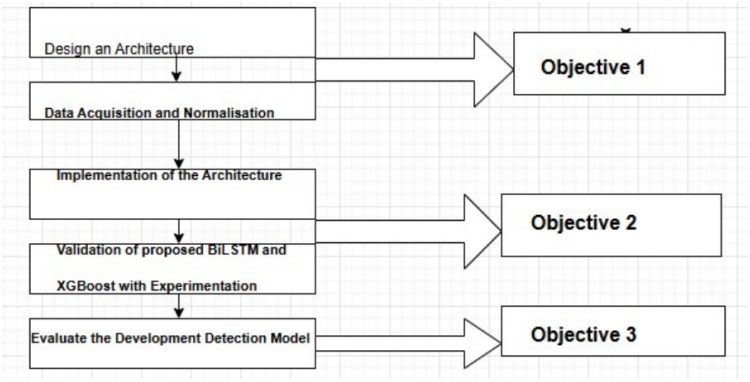


Figure 1. Methodology of BiLSTM and XGBoost for Cyberbullying Detection.

The first dataset was derived from a well-established Twitter benchmark corpus (currently known as X) and was designed to distinguish between offensive and neutral online messages [38] (p. 5). This collection comprised roughly 37,373 text samples, each annotated with a binary label 1 for offensive content and 0 for non-offensive posts. The structural layout and functional components of the implemented BiLSTM model are illustrated in **Figure 2**, which outlines the configuration utilized throughout this study.

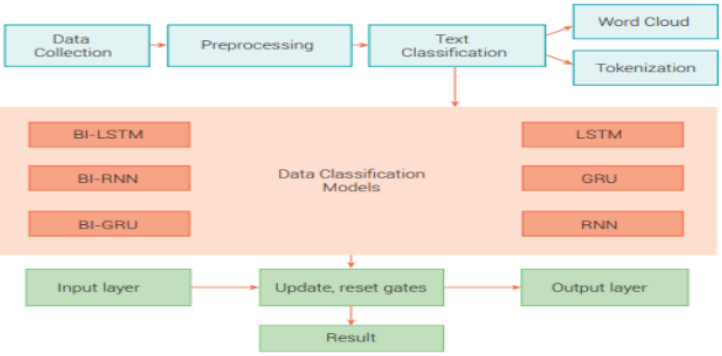


Figure 2. The Architecture of BiLSTM for Cyberbullying Detection on Social Media [18] (p. 7).

The second dataset was obtained from Facebook groups and community pages containing comments and discussions related to social behavior [39] (p. 5). It included roughly 20,000 records, focusing on identifying linguistic indicators of cyberbullying such as hate speech, racism, discrimination and verbal aggression. The data labeling process was conducted manually: posts with abusive or harmful expressions were assigned the label 1, while neutral or non-harmful entries were assigned 0. Out of these, approximately 12,000 were negative (offensive or harmful) and 8,000 were neutral or positive. **Figure 3** describes the distribution of short tweets containing fewer than ten words. The combined dataset was restricted to English-language text. Before training, data pre-processing and cleaning were conducted to remove duplicates, hyperlinks, and special symbols. Tokenization and lowercasing ensured uniformity, while stop-word removal and lemmatization minimized redundancy.

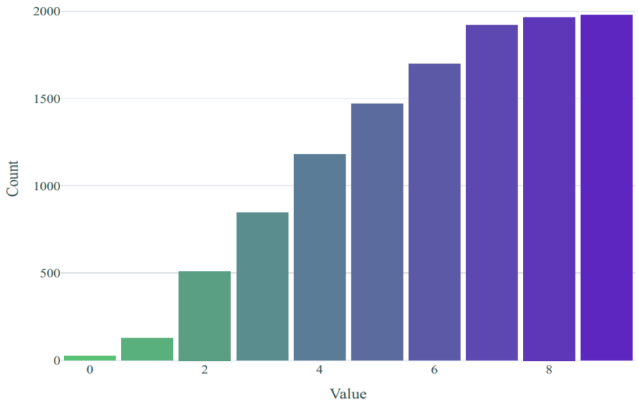


Figure 3. Distribution of Short Tweets Containing Fewer than Ten Words (Tweet Length Analysis).

Figures 3 and **4** visualize word-length distributions, showing that most tweets contained between 9 and 11 words, while the longest entry contained 52 words. **Table 1** displays representative samples of offensive and non-offensive tweets used for model evaluation. A sample of cyberbullying tweets from the Twitter dataset is displayed in **Table 1**.

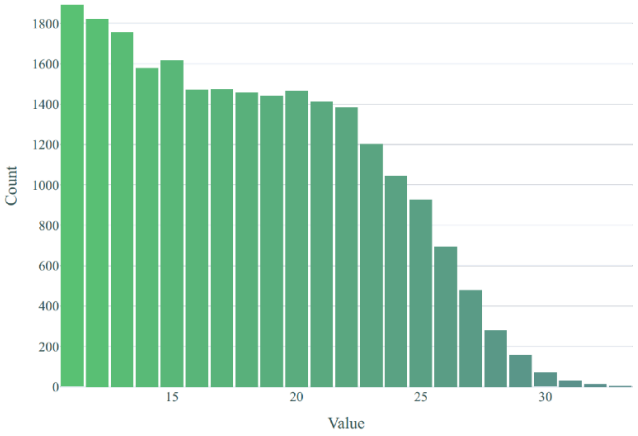


Figure 4. Word Count Distribution for Longer Tweets within the Dataset.

Table 1. Representative Examples of Cyberbullying-Related Posts Extracted from the X (Twitter).

ID	Sample Tweet (Cyberbullying Content)	Predicted Category	Label
1	Individuals who are overweight are useless.	Offensive (online harassment)	1
2	Are you referring to men? No, the man is simply gay and not a manager.	Offensive (online harassment)	1
3	Similar to shadows, fake pals remain in your life at your best times but vanish during your worst.	Non-Offensive (online harassment)	0
4	You're a large black person.	Offensive (online harassment)	1
5	Something is so cool today. That same item is garbage tomorrow. It won't matter next month.	Non-Offensive (online harassment)	0

3.2. Data Preprocessing

Each dataset underwent text normalization to ensure consistent input for model training. The preprocessing steps included:

(1) Conversion of all text to lowercase.
(2) Removal of punctuation, emojis, and hyperlinks.
(3) Application of tokenization using the NLTK library.
(4) Lemmatization to standardize word forms.
(5) Filtering of tokens to eliminate non-alphabetic strings.
(6) This step minimized noise within the dataset and enhanced the precision of feature extraction. After cleaning, the data were partitioned into three subsets training, validation, and testing using an 80:10:10 proportional split to enable effective cross-validation during model evaluation.

3.3. Embedding Layer

To effectively represent linguistic meaning and contextual relationships among words, this study adopted the Word2Vec embedding model, which converts textual data into compact, continuous-valued vector spaces. Within this representation, words sharing similar semantics are positioned close to one another in the vector space. Two primary Word2Vec training paradigms were employed in this study: the context-prediction model (commonly known as Continuous Bag-of-Words, or CBOW) and the target-prediction model (Skip-gram). The CBOW configuration predicts a central word based on the surrounding context, whereas the Skip-gram variant performs the inverse operation inferring nearby words from a specified target term.

For this research, the CBOW method was selected as the principal embedding approach because of its effectiveness in capturing contextual correlations between offensive and non-offensive expressions within social media text. To accelerate the training phase, the Hierarchical SoftMax optimization technique was implemented. This algorithm structures the vocabulary into a Huffman tree, thereby improving computational efficiency during probability estimation. The hidden-layer output for each token was obtained by averaging the vectors of its contextual words, as represented in Equation (1).

Equations (2)–(4) describe the probabilistic and loss functions used to fine-tune the model parameters. The training objective centered on maximizing the probability of accurately predicting semantically related words while concurrently minimizing cross-entropy loss, ensuring efficient convergence and stable learning performance.

Figure 5 illustrates the operational structure of the Word2Vec embedding model, which utilizes two key training paradigms to learn word relationships within a textual dataset the context-based CBOW model and the target-prediction Skip-gram model. Each serves a complementary linguistic role: in CBOW, the model infers a central term from its neighboring words, whereas Skip-gram functions inversely by estimating surrounding words from a specified target. Given their conceptual similarity, both approaches share comparable mathematical foundations, which are analyzed collectively in this study.

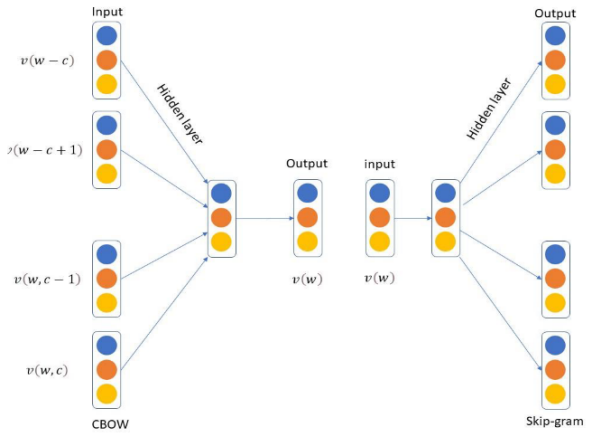


Figure 5. Distribution of Tweets with Higher Word Counts (Word2Vec Representation Overview).

To improve training efficiency and computational scalability, this work adopts optimization methods including Hierarchical SoftMax and negative sampling. The Hierarchical SoftMax structure organizes the vocabulary into a Huffman tree ordered by word occurrence frequency, with each internal node corresponding to a probabilistic output path. In this hierarchical setup, the traditional dense output layer is replaced with a more efficient tree-based structure, significantly reducing computational overhead during probability estimation. Consequently, the hidden-layer representation is formulated as the averaged vector of contextual embeddings, as expressed in Equation (1).

$$h = \frac{1}{c} \sum_{u \in context(w)} v(u) \tag{1}$$

Here, the term $context(w)$ denotes the group of words surrounding the target token w, which collectively provide its semantic background. The symbol $v(u)$ refers to the embedding vector corresponding to a particular contextual word u, while c represents the total number of words included in the context window for w. Using these notations, Equation (2) expresses the conditional probability of observing the target word w given its context, as derived from these definitions.

$$P\left(w \middle| context(w)\right) = \exp\left(u\_w \cdot v^1\right) / \sum\_\{u \in v\} exp(v\_u \cdot v^1(context(w))) \tag{2}$$

Let (w) denote the j-th internal node encountered along the path that connects the root node to the target word "w" within the Huffman tree structure. The path length corresponding to the word w is represented as $(w) - 1$, while the vector term refers to the embedding linked with the respective internal node n on that path. The operator $\| \cdot \|$ defines a specific function, mathematically formulated in Equation (3).

$$p\left(w \middle| context(w)\right)) = \prod\nolimits_{j=1}^{\int (w)-1} \sigma(sign(w,j) \cdot v(n(w \cdot j)) \cdot v^1(context(w))) \tag{3}$$

In this formulation, the j-th binary digit within the Huffman encoding sequence corresponding to the target word w is denoted as $dwj+1$. For the given context window (representing either the surrounding or target term), the model optimizes its parameters by maximizing the likelihood function derived from the equation during the training process illustrated in **Figure 6**. The resulting loss expression, shown in Equation (4), represents the logarithmic form of this likelihood, which serves as the objective function for the model's learning process.

$$1 = logP\left(w \middle| context(w)\right) \tag{4}$$

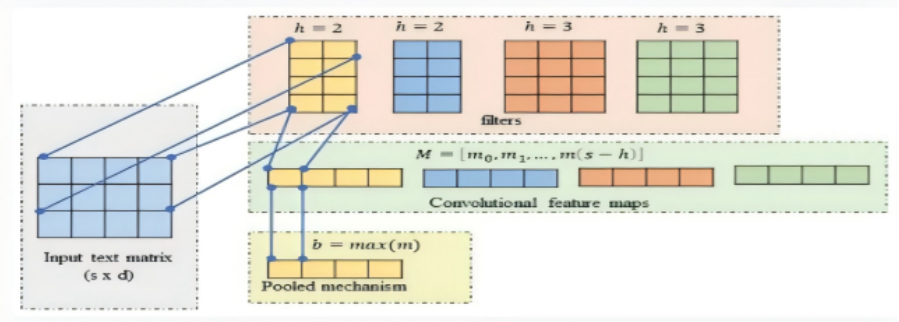


Figure 6. Mathematical Expression Applied during Model Optimization.

Equation (5) was used to derive the derivative l as a loss function with respect to the vector of the inner point (w).

$$\frac{\partial L}{\partial v(n(w,j))} = (\sigma(z_J) - 1 \cdot sign\ (w,J) \cdot v(context(w))) \tag{5}$$

Equation (6) expresses the gradient of the loss term (L) computed with respect to the embedding vector corresponding to each contextual word u within the model. Here, j = 1, 2, …, $l(w)$ − 1, where $l(w)$ represents the hierarchical depth associated with the word w.

$$\frac{\partial L}{\partial v(u)} = \frac{1}{c}\prod_{j=1}^{l(w)-1}\left(\sigma(z_j) - 1\right) \cdot \text{sign}(w, j) \cdot v(n(w, j)) \tag{6}$$

The Skip-gram and CBOW approaches operate as mutually reinforcing learning paradigms, each designed to achieve the same representational objective but from contrasting perspectives. In the Skip-gram configuration, the model learns outwardly predicting neighboring words that occur within a specific window around a target term, thereby modeling contextual expansion. Conversely, the CBOW framework performs the inverse process: it infers the central or target word based on the surrounding linguistic context, emphasizing inward contextual interpretation.

Within the framework of this study, the CBOW technique demonstrated particular effectiveness for cyberbullying detection, primarily because it mitigates issues related to data sparsity and enhances the model's ability to capture linguistic dependencies. By encoding subtle semantic relationships and contextual word interactions, the CBOW-derived embeddings allow the system to precisely recognize and categorize language patterns that reflect offensive or harmful online behavior.

3.4. Deep Neural Networks (DNN) Baseline Models

In this study, four distinct deep neural network (DNN) architectures, Conv1DLSTM, BiLSTM, LSTM, and CNN were developed and assessed to evaluate their effectiveness in detecting cyberbullying content. Each model was trained and tested using the same Twitter dataset to ensure uniformity and fairness in comparison. The subsequent subsections provide concise explanations of the design principles, training configurations, and operational steps applied in constructing these baseline DNN frameworks for cyberbullying classification.

3.4.1. Long Short-Term Memory and Bidirectional Long Short-Term Memory

The Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) models are enhanced variants of the Recurrent Neural Network (RNN) architecture, purpose-built for learning from data sequences where earlier inputs influence later outputs. Such sequential dependencies are typical in natural language processing, speech analysis, and other time-series data.

Unlike traditional RNNs that process information in a single temporal direction, the BiLSTM structure reads sequences in both forward and backward order. This bidirectional processing enables the network to access contextual information from both preceding and subsequent terms, thereby improving its ability to interpret meaning and achieve higher predictive precision.

One major limitation of conventional RNNs is the vanishing gradient problem, where long-term dependencies fade as gradient values diminish during training. The LSTM addresses this challenge through a specialized memory cell that dynamically manages information flow across time steps. It operates using three control gates:

i. The input gate, which filters and admits new information;
ii. The forget gate, which removes data that are no longer relevant;
iii. The output gate, which determines which processed information is passed to the next layer.

By managing these gates collaboratively, LSTMs can retain critical long-term dependencies while discarding irrelevant ones. This controlled mechanism allows the model to learn context-rich representations within sequential data, making it particularly effective for text classification tasks such as cyberbullying detection, where understanding the relationship between words across context is essential.

$$a^t = g(w_a[a^{t-1} - X^t] + b_q) \tag{7}$$

$$y^t = f(w_{y \cdot a^t} + b_y) \tag{8}$$

3.4.2. Convolutional Neural Network

At this computational stage, the model begins to extract structured feature representations from the input data, generating what are known as feature maps. The Convolutional Neural Network (CNN) architecture operates through multiple stacked convolutional layers, where each layer focuses on identifying specific localized attributes from the preceding representation.

When applied to Natural Language Processing (NLP) contexts, these convolutional layers are particularly effective in detecting short-term word associations and recurring semantic patterns across text sequences. This makes CNNs highly suitable for analyzing linguistic data characterized by local dependencies, such as phrases or n-grams that convey offensive or bullying intent.

In this study, convolutional operations were carried out on the output of the attention mechanism, using a linear filtering process to extract the most discriminative linguistic signals. The filter, denoted as f of size $e \times h \times h$ and iteratively slid across the embedded input matrix. Each social media post, represented as a sequence X containing x tokens, was first converted into an embedding vector of dimension e.

The convolutional transformation defined in Equation (9) then generated the feature map, summarizing critical textual and contextual characteristics for higher-level layers of the model, enabling improved pattern recognition and classification accuracy. The result of this generate a feature map $M = [m1, m2, ..., mx - h] = 1, 2,'''', -h$ as formulated in Equation (9).

$$m_t = f \times x_{t:i+h-1} \tag{9}$$

After the convolutional transformation, the resulting feature maps undergo a pooling (or sub-sampling) operation designed to condense the learned representations by reducing dimensionality while maintaining key information. Among several pooling approaches, max-pooling was adopted in this study for its effectiveness in highlighting the most influential linguistic and contextual patterns while suppressing redundant or less informative signals.

Let $i = 0, 1, ..., x - h$, and let $X_{i:j}$ denote a localized sub-matrix of X that spans the interval between indices i and j. As described in Equation (10), the max-pooling process selects the largest activation value, b, from each subregion of the feature map. This operation ensures that only the most prominent semantic cues are carried forward to deeper layers of the neural model, improving computational efficiency and helping the network focus on the strongest textual signals related to cyberbullying patterns.

$$B = \max_{0 \le t \le x-h} \tag{10}$$

The representations obtained from the pooling layer were merged to generate a composite feature vector, serving as a condensed depiction of the most salient linguistic and contextual properties identified during the convolutional stage. This aggregated vector was subsequently passed into the Fully Connected Layer (FCL), where higher-level abstractions were derived, enabling the model to interpret and categorize the extracted features effectively. This stage of computation transforms the localized textual representations into a global understanding suitable for classification tasks. The hierarchical progression of this process from convolution through pooling to full connection is conceptually visualized in **Figure 5**, demonstrating how features are incrementally refined and integrated to support accurate final predictions and informed decision-making within the model.

3.4.3. Fully Connected Layer (FCL)

The Fully Connected Layer (FCL) operates as the concluding stage of the neural architecture, where all previously extracted features are integrated and interpreted to produce the final classification output. In this stage, the pooled and concatenated feature representations are mapped through a dense network of interconnected neurons, allowing the system to learn complex decision boundaries that distinguish between cyberbullying and non-cyberbullying expressions.

The FCL applies a series of weighted transformations to project the high-dimensional feature vectors into a smaller decision space, thereby converting abstract contextual representations into concrete class predictions. This mechanism enables the model to consolidate all semantic cues derived from earlier convolutional and recurrent layers into a unified classification result, ensuring reliable detection performance.

$$H_t = SOftMax(w_t\, h_{t-1} + b_t) \tag{11}$$

where w_t and b_t are parameters learned in training, H_t is the obtained from the pooled concatenated feature vector and h_{t-1}^{h-1} is the feature map received from the CNN layers. The output layer performs the correct classification using the SoftMax function, as in **Figure 1**. The cross-entropy loss was minimized to learn the model parameters as the training objective using the Adam optimization algorithm. It is provided by Equation (12). At the final stage, the output layer illustrated in **Figure 1** applies the SoftMax activation mechanism, which transforms the raw network outputs into normalized probability scores corresponding to each target class. Training was conducted using the Adam optimizer, an adaptive learning algorithm that dynamically adjusts parameter updates to accelerate convergence and stabilize learning. This process systematically tunes the model's weights and biases to minimize the cross-entropy objective function, thereby enhancing both accuracy and generalization.

$$CrossEntropy(p, q) = -\sum_x p(x) log q(x) \tag{12}$$

Here, p denotes the true class distribution, and q represents the SoftMax-predicted probabilities. The resulting negative log-likelihood quantifies the divergence between predicted and actual outcomes. In this formulation, p is encoded as a one-hot vector, where each element corresponds to a distinct token or character within the analyzed social media corpus.

3.5. Stacked Ensemble Workflow for Cyberbullying Detection

Cyberbullying datasets are gathered from social media platforms and assembled into a tagged corpus that includes text samples of both cyberbullying and non-cyberbullying. The text is normalized, which includes removing punctuation, special characters, URLs, and lowercasing. Lemmatization and stop-word elimination are used to increase linguistic consistency and lower noise. Using Word2Vec with the Continuous Bag-of-Words (CBOW) architecture, the cleaned text is tokenized and transformed into numerical representations, producing dense semantic word embeddings. To guarantee uniform input dimensions for the neural network models, the tokenized text sequences are either padded or shortened to a predetermined sequence length. Local n-gram features are captured by CNN, long-term dependencies are captured by LSTM, bidirectional contextual information is captured by BiLSTM, and convolutional feature extraction and sequential learning are combined by Conv1D-LSTM.

Following training, a Softmax output layer is used by each base learner to generate probability ratings for every class. For every sample, a meta-feature vector is created by concatenating the probability outputs from all base learners. The basic models' combined predictions are represented by this vector. An Extreme Gradient Boosting (XGBoost) classifier receives the meta-feature matrix as input. In order to integrate the base learners' predictions, XGBoost learns the best weights and decision bounds. The same pre-processing and embedding procedures are applied to fresh text data during inference. The final classification result is obtained by combining and feeding the prediction probabilities generated by the base models into the trained XGBoost meta-learner. The final stacked model is evaluated using metrics such as accuracy, precision, recall, and F1-score to assess classification performance and compare it with individual base models. **Algorithm 1** stacked ensemble workflow for cyberbullying detection algorithm.

Algorithm 1 Stacked Ensemble Workflow for Cyberbullying Detection Algorithm

Input: Dataset D = {(x_i, y_i)}_{i=1..N}, where x_i represents a social media text instance and y_i ∈ {0,1} denotes the class label (cyberbullying or non-cyberbullying). Output: Final cyberbullying prediction ŷ.
Step 1: Data Preprocessing: Convert all text to lowercase, Remove URLs, punctuation, numbers, and special characters, Tokenize text into words, Remove stopwords. Apply lemmatization to normalize tokens.
Step 2: Feature Representation: Train a CBOW Word2Vec embedding model on the corpus, Transform each processed text x_i into a sequence of embedding vectors, Pad sequences to a fixed maximum length.
Step 3: Base Model Training: Train the following deep learning base learners using the embedding sequences: CNN, LSTM, BiLSTM, and Conv1D–LSTM hybrid model, For each base model M_j, generate class probability predictions p_j(x_i) = [p_j0, p_j1].
Step 4: Meta-Feature Construction (Stacking Layer): Concatenate prediction probabilities from all base models to create a meta-feature vector: z_i = [p1(x_i), p2(x_i), p3(x_i), p4(x_i)].
Step 5: Meta-Learner Training: Train an XGBoost classifier using the meta-feature vectors z_i and labels y_i.
Step 6: Final Prediction: For a new input text x, apply preprocessing and embedding steps, Generate probability outputs from all base models, Construct the meta-feature vector z, Use the trained XGBoost meta-learner to produce the final prediction ŷ.
Step 7: Evaluation
Evaluate model performance using Accuracy, Precision, Recall, and F1-Score on the test dataset.

4. Results

4.1. Experimental Setup

The experimental phase of this research was conducted on Google Colab, leveraging GPU acceleration and Python version 3.8 to enhance computational performance. Both the proposed hybrid cyberbullying detection framework and the baseline deep learning models were developed using a suite of Natural Language Processing (NLP) tools within the TensorFlow library, which also provides integrated support for computer vision operations. The model architecture was intentionally optimized to maintain computational efficiency by eliminating redundant hidden nodes and fine-tuning critical hyperparameters within the dense network layers.

Text preprocessing and tokenization were performed in TensorFlow, where each text sample was converted into a structured sequence matrix comprising 35,873 tweets. This transformation segmented raw textual data into discrete tokens, enabling the model to capture contextual dependencies and extract semantically rich features from the input corpus.

Prior to tokenization, a comprehensive data-cleaning pipeline was applied to remove duplicates, incomplete records, irregular text patterns, and missing entries. To reduce noise and improve representation quality, non-informative stop words that contributed minimally to the overall meaning of sentences were excluded from the dataset. This procedure facilitated a reduction in feature dimensionality and enhanced the overall signal-to-noise ratio during representation learning. A comprehensive overview of the model configuration, experimental setup, and layer-specific parameters utilized in this study is summarized in **Table 2**.

Table 2. Experimental Parameters and Layers of the Proposed Model.

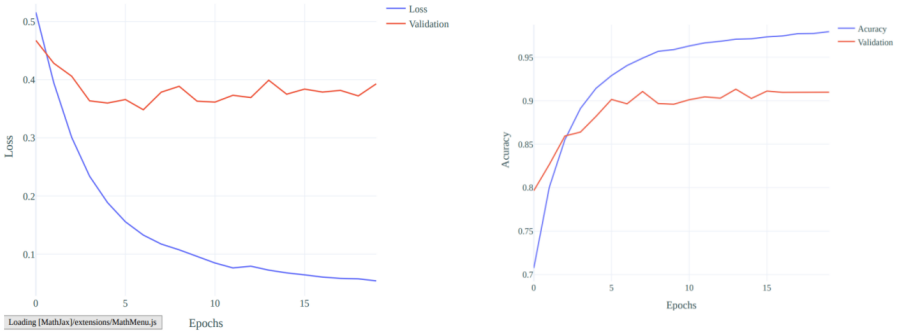
Layers	Layer Name	Kernel × Unit	Other Parameters
1	Conv1D	72 × 128	Activation = ReLU, Strides = 3
2	Batch Norm	-	-
3	Global Max Pool	-	Stride = 3
4	Conv1D	-	Activation = ReLU, Strides = 3
5	Batch Norm	-	-
6	Max Pool	-	Pool Size = 2, Stride = 2
7	Conv1D	3 × 512	Activation = ReLU, Stride = 1
8	Conv1D	3 × 128	Activation = ReLU, Stride = 1
9	Flatten	-	-
10	Dense	1 × 512	-
11	Dense	2	Activation = SoftMax

The embedding layer of the proposed model employed a hybrid initialization scheme that merged Word2Vec and Continuous Bag-of-Words (CBOW) representations to strengthen contextual encoding. Word embeddings were generated from 233 distinct terms obtained from the mixed dataset and 149 tokens from the Twitter-specific corpus, resulting in an 87-dimensional vector space. Within the deep learning structure, each neuron operated with 32–256 memory units, scaled in increments of 32.

Through extensive experimentation, the stacked hybrid network attained its best convergence behavior when trained using the Adam optimizer implemented in the TensorFlow environment. To maintain computational efficiency, the number of iterations was deliberately limited, allowing the framework to perform automatic hyperparameter refinement. A regularization rate of 0.25 was consistently applied during training to mitigate overfitting, with five to ten training runs typically required to achieve stable outcomes—averaging two or three successful trials per configuration.

For filter optimization, convolutional parameters were tuned between 32 and 132 filters, with kernel widths of 2 and 4 yielding the best results. The Fully Connected Layer (FCL) dimension was fixed at 132, while embedding weights were initialized using the Glorot uniform distribution. The complete model underwent training for 20 epochs under the Adam optimization strategy. Empirical tests on the Twitter dataset confirmed that a batch size of 42 produced the most consistent convergence, particularly for sequences exceeding 10–20 words. The dropout ratio remained constant at 0.25 throughout all runs, while the learning rate was dynamically adjusted between 0.001 and 0.1 to preserve performance stability. The SoftMax function was retained for final classification, and all other structural components followed the same architectural logic, with slight adjustments applied when LSTM or BiLSTM layers replaced Conv1D. **Figure 7** shows Evaluation results showing the training loss trend and validation

accuracy of the proposed stacked ensemble framework applied to the Twitter dataset.

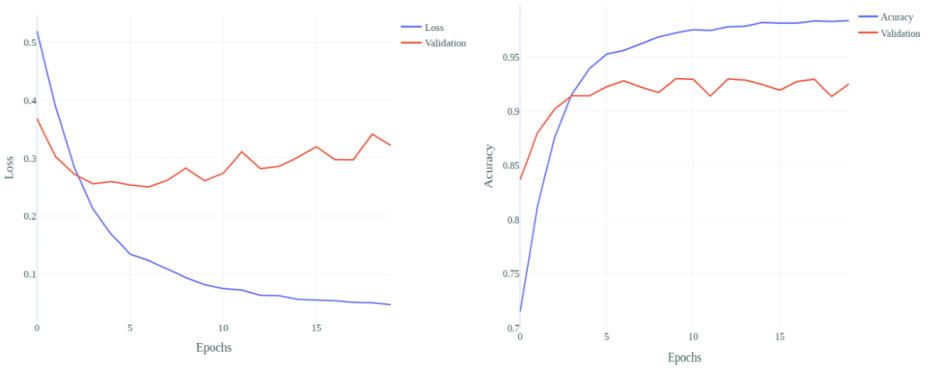


(**a**) Proposed stacked ensemble learning model loss. (**b**) Proposed stacked ensemble learning model testing accuracy.

Figure 7. Evaluation results showing the training loss trend and validation accuracy of the proposed stacked ensemble framework applied to the Twitter dataset.

In addition to the primary experiment, an extended evaluation was carried out using a broader dataset that combined entries from both Twitter and Facebook communities. This supplementary test aimed to further validate the resilience and generalization ability of the proposed stacked ensemble framework across multiple social media contexts, aligning with dataset variations explored in prior studies. The aggregated dataset encompassed diverse forms of cyberbullying-related expressions, including racial slurs, discriminatory comments, offensive language, and verbal aggression.

The training configuration and hyperparameter settings were kept consistent with the earlier experimental setup, with each model trained for 20 epochs to enable uniform comparison. The relationship between model loss and validation accuracy observed for the stacked ensemble is depicted in **Figure 8a**, while **Figure 8b** illustrates the corresponding accuracy performance trends during the validation phase. The subsequent sections present a comprehensive discussion of the outcomes obtained from the baseline deep-learning models, the proposed ensemble configuration, and comparative analyses involving the standard BERT and its fine-tuned counterpart.



(**a**) Proposed stacked ensemble learning model loss. (**b**) Proposed stacked ensemble learning model testing accuracy.

Figure 8. Variation of training loss and validation accuracy for the proposed stacked ensemble framework, evaluated on the combined social media dataset encompassing Twitter and Facebook sources.

4.1.1. Performance Matrix

i. To evaluate the predictive performance of the developed models, four major assessment criteria were utilized: accuracy, precision, recall, and F1-score. Together, these measures offer an integrated perspective on the model's ability to correctly distinguish between cyberbullying and non-cyberbullying posts. Each metric captures a unique performance dimension, supporting a balanced assessment of detection reliability, sensitivity, and consistency.

ii. Within the domain of cyberbullying identification, accuracy reflects the overall proportion of correctly classified instances, encompassing both harmful and non-harmful tweets relative to the complete dataset. It serves as a general indicator of the model's overall effectiveness. Mathematically, accuracy represents the ratio of correctly predicted samples to the total number of evaluated records, as defined in Equation (13).

$$ACCURACY = \frac{(tp + tn)}{(tp + fp + tn + fn)} \tag{13}$$

iii. Precision represents the model's capacity to accurately identify genuine positive samples from all instances it has marked as positive. In practical terms, it measures the reliability of the model's positive predictions by determining the proportion of detected tweets that are truly cyberbullying-related.

iv. Recall—sometimes referred to as sensitivity—indicates the percentage of actual positive instances correctly identified by the model. It demonstrates how effectively the system can recover all relevant bullying posts from the dataset, minimizing the number of missed or overlooked cases.

v. The F1-score combines both precision and recall into a unified metric by computing their harmonic mean, thus balancing prediction accuracy and completeness. This metric is particularly valuable for imbalanced datasets, where one class (for example, non-bullying tweets) significantly outnumbers another, helping to ensure fairer evaluation across categories.

Finally, these evaluation parameters were comprehensively applied to gauge the performance and robustness of the proposed cyberbullying detection framework. Their mathematical formulations are provided in Equations (14)–(16).

$$Precision = \frac{tp}{(tp + fp)} \tag{14}$$

$$Recall = \frac{tp}{(tp + fn)} \tag{15}$$

$$F1\ score = \frac{2 \times precision \times recall}{recision + recall} \tag{16}$$

vi. In binary classification problems, such as cyberbullying identification, the Precision–Recall (PR) curve provides a reliable means of evaluating a model's predictive effectiveness, especially in stacked ensemble systems. This graphical representation depicts how precision and recall vary at different decision thresholds, illustrating the balance between these two measures. In this study, recall refers to the proportion of correctly identified positive examples among all genuine positive cases, whereas precision indicates the share of accurately predicted positive samples relative to all instances classified as positive. Because both precision and recall account for false positives and false negatives, their combination produces a comprehensive view of the model's detection accuracy. Plotting precision against recall at multiple threshold values yields the PR curve, from which the Area Under the Precision–Recall Curve (AUPRC) is derived. A higher AUPRC value reflects stronger discriminative capability and classification stability. In application, precision demonstrates how closely the system's positive predictions align with the actual occurrences of cyberbullying, while recall represents the extent to which the model captures all relevant bullying instances. A well-optimized classifier achieves both high precision (fewer false positives) and high recall (fewer false negatives). As the PR curve illustrates, the ratio of correctly predicted positives to all identified positives acts as an indicator of model robustness. Classifiers whose PR points approach the ideal coordinate (1, 1) exhibit excellent performance, whereas those near zero display weaker detection accuracy. The precision–recall distribution for the proposed stacked ensemble framework in detecting online harassment is depicted in **Figure 9**.

Beyond the Precision–Recall (PR) evaluation, this study also applied the Receiver Operating Characteristic (ROC) analysis as a complementary assessment tool to gain deeper insight into model performance. The ROC curve visualizes the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across a range

of classification thresholds, offering a two-dimensional perspective on the system's discriminative ability. This visualization enables a more detailed interpretation of how efficiently the proposed model separates cyberbullying-related posts from non-bullying ones under varying decision conditions.

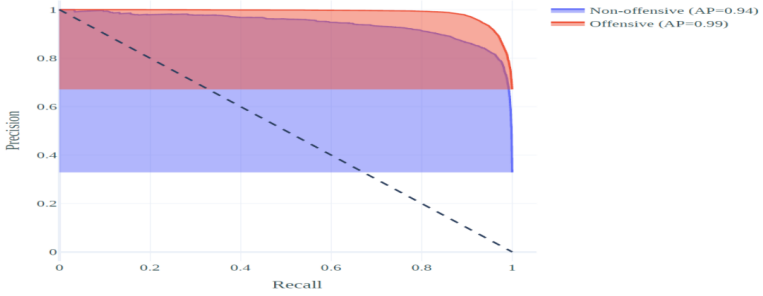
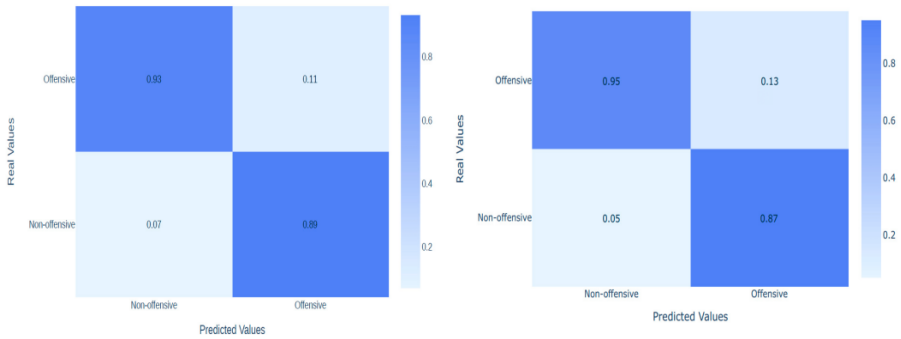


Figure 9. Precision–Recall Curve for Stacked Ensemble Model.

To maintain fairness and avoid performance bias, the model was evaluated using a distinct test dataset that had not been involved in the training stage. This procedure ensured that all deep learning (DL) models were compared under uniform experimental conditions. The confusion matrices for the two best-performing architectures are presented in **Figure 10**, illustrating a four-component breakdown that includes true positives, true negatives, false positives, and false negatives. Together, these elements provide a comprehensive visualization of model accuracy and reveal the classifier's overall prediction behavior.



(**a**) Confusion matrix of stacked ensemble model (**b**) Confusion matrix of M-BERT-based model

Figure 10. Confusion Matrix of DL Predictor's Models.

i. True Positive (TP): This occurs when the system correctly recognizes and labels a tweet containing harmful or abusive expressions as offensive.
ii. False Positive (FP): This case arises when the model mistakenly classifies a harmless or neutral tweet as offensive, resulting in an incorrect detection of cyberbullying.
iii. False Negative (FN): This outcome takes place when the model fails to identify bullying content that is genuinely harmful, thus labeling it as non-offensive.
iv. True Negative (TN): This represents a correct prediction where the model accurately detects that a tweet is non-harmful and assigns it a non-offensive label.
v. The Area Under the Curve (AUC) metric was utilized as a principal measure of the binary classification performance of the developed models, particularly for cyberbullying detection across social media datasets. The AUC quantifies how effectively the classifier distinguishes positive (cyberbullying) instances from negative (non-cyberbullying) ones. It does this by visualizing the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across multiple threshold levels.
vi. In this context, TPR corresponds to the proportion of harmful tweets correctly identified as abusive, while FPR denotes the percentage of neutral tweets that were incorrectly marked as bullying. A higher AUC value therefore signifies stronger classification consistency and a more robust discriminatory capability.

4.1.2. Performance Outcomes of Baseline Architectures

The experimental findings for the baseline deep learning frameworks are outlined in **Table 3**. Four core architectures, Conv1D-LSTM, BiLSTM, LSTM, and CNN were designed and evaluated using Word2Vec vector representations generated from the Continuous Bag-of-Words (CBOW) feature extraction approach, applied to the Twitter dataset. The architectural structures and tuning parameters for each model are presented in **Table 4**. Among the baseline models, the Conv1D-LSTM framework demonstrated the highest performance, achieving an accuracy of 0.8649, precision of 0.8142, recall of 0.7281, and an F1-score of 0.8281. These results reflect its capacity to capture both short- and long-range contextual dependencies in textual data. In contrast, the BiLSTM model yielded comparatively lower outcomes, with an accuracy of 0.7795, precision of 0.8373, recall of 0.8130, and an F1-score of 0.8041, indicating minor limitations in identifying cyberbullying-related language patterns. To enhance predictive reliability, a stacked ensemble configuration was developed by combining the discriminative capabilities of all four baseline models. This integrated framework produced a substantial gain in classification accuracy, surpassing both the baseline BERT (0.921) and the fine-tuned BERT (0.9384) implementations. Overall, the proposed hybrid stacked-ensemble system attained an accuracy of 0.974, demonstrating superior detection precision and the most balanced performance profile among all compared architectures.

Table 3. Comparison Analysis between Baseline and Models on the Twitter Dataset.

No.	Algorithm	Accuracy (%)	Precision	Recall	F1-Score
1	LSTM	0.8011	0.8142	0.7281	0.8281
2	Conv1DLSTM	0.8649	0.8146	0.8919	0.8317
3	CNN	0.8496	0.8836	0.7908	0.8720
4	BiLSTM	0.7795	0.8373	0.8130	0.8041
5	BERT	0.921	0.915	0.915	0.9149
6	Tuned-BERT	0.9384	0.92	0.91	0.92
7	Stacked	0.974	0.950	0.92	0.964

Table 4. Comparison Analysis between Baseline and Models on Facebook Dataset.

No	Algorithm	Accuracy (%)	Precision	Recall	F1-Score
1	BERT	0.9042	0.9051	0.9034	0.9043
2	Tuned-BERT	0.9198	0.9262	0.9123	0.9191
3	Stacked	0.9097	0.9122	0.9082	0.9102

The overall evaluation of the implemented deep learning frameworks including the stacked ensemble, baseline BERT, and the fine-tuned BERT configurations is illustrated in **Figure 11**. The proposed stacked ensemble framework, which integrates outputs from multiple pre-trained neural architectures, exhibited superior predictive capability compared to the standalone BERT-based systems. This improvement demonstrates the benefit of combining diverse model architectures to capture complementary linguistic and contextual representations of social media text.

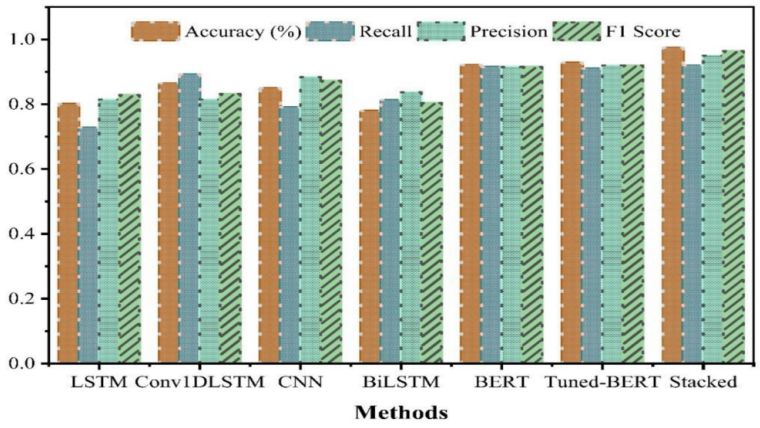


Figure 11. Comparative Summary of the Proposed Deep Learning Models for Cyberbullying Detection.

The fine-tuned BERT variant, optimized through additional domain-specific retraining, further enhanced the model's ability to recognize cyberbullying-related expressions. Its adaptation to social-media-specific patterns and language styles contributed to greater contextual sensitivity in classification tasks.

In comparative terms, the stacked-ensemble configuration achieved the highest level of performance across all evaluation parameters, recording an accuracy of 0.974, precision of 0.950, recall of 0.920, and an F1-score of 0.964. These outcomes validate the effectiveness of combining multiple deep-learning networks within a unified system, leading to more reliable classification outcomes and improved detection of subtle or implicit online aggression.

The evaluation compared the performance of the stacked-ensemble framework with the baseline architectures, using a combined dataset drawn from two major social media sources Facebook and X (formerly Twitter). This hybrid dataset provided a more realistic representation of online user interactions and linguistic diversity. A detailed performance comparison, including accuracy, precision, recall, and F1-score metrics, is presented in **Table 5**.

Table 5. Comparison of Models' Complexity and Statistical Analysis.

No	Algorithm	Accuracy (%)	Time Complexity
1	BERT baseline	92.1%	1 h 6 min
2	Modified-BERT	93.84%	1 h 2 min
3	Proposed Stacked	97.4%	3 min 9 s

Among the assessed models, the fine-tuned BERT recorded the highest overall detection performance, achieving an accuracy of 91.98%, with consistently high precision, recall, and F1-score values. Although its accuracy was marginally higher than both the baseline BERT and the stacked-ensemble model, the fine-tuned variant required a substantially longer training duration, approximately 41 min and 23 s per cycle. In contrast, the stacked-ensemble configuration achieved a comparable accuracy of 90.97% while completing its training in just 2 min and 45 s. This demonstrates the superior computational efficiency of the ensemble system, making it particularly well-suited for large-scale or real-time cyberbullying detection tasks where rapid deployment and reduced training time are critical.

4.1.3. Comparison of Proposed Models' Complexity and Statistical Analysis

In this study, three deep learning classifiers were developed utilizing a Word2Vec embedding framework with a Continuous Bag-of-Words (CBOW) representation for feature generation. A comparative evaluation was carried out to examine the models' effectiveness in terms of classification accuracy and computational efficiency. The outcomes of this evaluation, which illustrate the balance between predictive capability and processing time, are summarized in **Table 5**.

The baseline model, identified as the BERT architecture, recorded an overall accuracy of 92.1%. BERT (Bidirectional Encoder Representations from Transformers) remains a widely recognized pre-trained transformer model known for its exceptional capability to capture contextual and semantic dependencies across a broad spectrum of natural language processing (NLP) tasks. Within the context of this experiment, the baseline configuration required approximately 1 h and 6 min to complete training and evaluation on the dataset.

An adapted version, referred to as Modified-BERT, achieved a slightly improved accuracy of 92.84%, suggesting that minor adjustments to model architecture or hyperparameters positively influenced its predictive capacity. Its runtime was moderately shorter, completing the training phase in 1 h and 2 min.

Across all experimental configurations, the Proposed Stacked-Ensemble model achieved the best overall result, attaining an accuracy of 97.4%. This highlights the benefit of integrating multiple neural network layers and diverse feature-learning strategies to enhance classification precision. Despite outperforming other models in accuracy, it also demonstrated exceptional computational efficiency completing its entire training process in only 3 min and 9 s.

Additional validation performed on a separate dataset (summarized in **Table 6**) confirmed both the scalability and efficiency of the ensemble approach. On this secondary dataset, the model achieved an accuracy of 90.97% with a total training time of just 2 min and 45 s, confirming its suitability for real-world, time-sensitive cyberbullying detection tasks. Although the Modified-BERT model produced a marginally higher score of 91.98%, it required

substantially longer processing time (41 min and 23 s), reinforcing the ensemble model's advantage in rapid, high-accuracy deployment scenarios.

Table 6. Model Complexity and Statistical Analysis on Facebook Dataset.

No	Algorithm	Accuracy (%)	Time Complexity
1	BERT baseline	90.42%	44 min 25 s
2	Modified-BERT	91.98%	41 min 23 s
3	Proposed Stacked	90.97%	3 min 45 s

4.1.4. Comparison with Literature

A comparative assessment was conducted on several machine learning and deep learning algorithms using the Twitter dataset to identify and classify instances of cyberbullying. The performance outcomes of these models are summarized in **Table 7**. Each algorithm was evaluated through core metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive measure of their detection capabilities.

Table 7. Comparison with Related Literature.

Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score
Twitter	Logistic Regression	90.57	0.951	0.905	0.928
	LGBM Classifier	90.55	0.9614	0.895	0.927
	Random Forest	89.8	0.933	0.913	0.923
	SVM	67.13	0.933	0.913	0.923
	Stacked (ours)	97.4	0.950	0.92	0.964

This analysis served as an objective benchmark, offering insight into how different computational approaches perform when distinguishing harmful content from non-offensive text. The outcomes displayed in **Table 7** align closely with those reported in earlier studies that utilized the same dataset, thereby reinforcing the consistency, methodological validity, and reliability of the comparative evaluation.

The results of this study demonstrate a notable improvement in cyberbullying detection across major social networking platforms, including Facebook and X (formerly Twitter). This emphasizes the importance of early recognition and timely response in reducing the psychological and social consequences of online harassment. The proposed system distinguishes itself from traditional models through its hybrid ensemble structure, which combines multiple feature extraction methods to deliver higher accuracy and superior computational performance.

Through the integration of several deep learning architectures, the framework enhances both robustness and predictive stability, making a significant contribution to the field of automated online abuse detection. The effectiveness of the system largely stems from its feature extraction process, which utilizes Continuous Bag-of-Words (CBOW) and Skip-gram methods to initialize embedding layers. This design allows the model to capture deep semantic and linguistic relationships within the dataset. Additionally, the incorporation of convolutional and pooling layers helps minimize feature redundancy while preserving contextual associations among textual elements. Consequently, this combined mechanism strengthens the model's capacity to accurately detect and categorize cyberbullying content, supporting reliable detection in real-world social media environments.

4.2. Discussion

4.2.1. Generalizability of Data

Although the proposed model demonstrates strong predictive performance, its ability to generalize effectively across different platforms and domains remains a potential concern. The performance of pre-trained embeddings is heavily influenced by the quality, contextual relevance, and domain compatibility of the training datasets. When the linguistic or structural characteristics of a new dataset differ considerably from the source corpus, the model's accuracy and contextual understanding may decline. Hence, it is critical that the training data reflect the language patterns and social contexts of the target environment to ensure dependable model performance across various applications.

4.2.2. Data Bias

The diversity and representativeness of training data significantly affect the reliability of cyberbullying detection systems. Datasets that exhibit bias such as the overrepresentation of specific cultures, languages, or user communities can hinder the model's ability to capture the full spectrum of online interactions. This limitation may lead to partial or skewed predictions. To overcome these issues, developers should integrate demographically varied samples, maintain balanced datasets, and continuously refine data composition. Reducing bias in this way strengthens the model's fairness, adaptability, and generalization across global social platforms.

4.2.3. Class Imbalance

While focal loss functions are designed to reduce the impact of class imbalance, they do not completely solve the problem especially in datasets with large discrepancies between positive and negative samples. A strong imbalance can cause models to overfit to dominant categories, making them less sensitive to minority classes. To enhance detection performance, complementary rebalancing approaches such as data augmentation, re-sampling, or hybrid balancing techniques should be adopted. These methods ensure more equitable representation of minority samples during training, thereby improving the model's recall, precision, and overall classification consistency.

4.2.4. Interpretability

Although deep learning frameworks achieve impressive predictive results, they often function as opaque or "black-box" systems, limiting insight into their decision-making processes. This lack of interpretability makes it challenging to identify which contextual or linguistic features drive a model's predictions. To address this, integrating explainability mechanisms such as attention-map visualization, layer-wise relevance analysis, or feature-importance mapping is strongly encouraged. These tools enhance transparency, improve understanding of the model's reasoning, and build user confidence in automated cyberbullying detection systems.

5. Conclusions and Future Work

In recent years, remarkable progress has been achieved in the field of cyberbullying detection, with ensemble-based deep learning models playing a pivotal role in enhancing classification reliability and robustness. The architectures explored in this research, CNN, Conv1D-LSTM, and LSTM demonstrated strong capability in identifying offensive and abusive language across social platforms such as Facebook and X (formerly Twitter). This study employed two independent datasets to comprehensively assess the performance of the stacked-ensemble model. Experimental results confirmed that the model was able to detect and categorize harmful expressions with high efficiency, achieving superior scores in accuracy, precision, recall, F1-measure, and detection time. These findings validate the proposed architecture as an effective approach for recognizing subtle and context-dependent forms of online harassment.

While the outcomes are promising, further refinement is essential to enhance adaptability, scalability, and cross-domain consistency. Future research should explore more adaptive feature-learning mechanisms and next-generation deep learning models that can better accommodate dataset variations and cultural diversity in language use. Integrating explainable AI methods would also strengthen interpretability, helping researchers and practitioners understand the decision-making behavior of detection systems. Ultimately, this research makes a meaningful contribution to the growing body of knowledge in automated cyberbullying detection, offering a scalable and reliable solution for improving digital safety. The proposed model provides a strong foundation upon which future work can build to advance real-time, ethical, and inclusive detection systems suited for the dynamic nature of modern social media communication.

The inherent architectural complexity of transformer models like BERT, which have substantially more parameters and self-attention layers than the lighter CNN–LSTM–BiLSTM ensemble used in this study, is primarily responsible for the observed discrepancy in training time. We acknowledge, however, that the training variables (such as batch size, sequence length, and optimization settings) were not perfectly aligned, which could have an impact on comparative fairness. Furthermore, due to limited hyperparameter tweaking and dataset peculiarities, the fine-tuned BERT performance (F1 = 0.92) is marginally lower than typical benchmark results. Additionally, the pre-treatment pipelines are different since BERT typically handles raw text, whereas the ensemble employs lemma-

tization and stop-word removal. Future work will ensure standardized training settings and optimized transformer fine-tuning for a more balanced comparison.

Author Contributions

Y.A.M. as the principal investigator, conceptualized the research idea and designed the study methodology. J.A.O. provided comprehensive academic supervision, ensuring the proper structuring, refinement, and scholarly alignment of the research work. F.O. served as the co-supervisor, while A.O.I. participated as the project assessor during each presentation and examination phase. A.U. and M.B. contributed to the initial draft development and manuscript preparation. P.C.A. supported the review, editing, corrections required and formatting of the document for journal publication. All authors have read and agreed to the published version of the manuscript.

Funding

The research was self-sponsored by Y.A.M., who personally financed all aspects of the study.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

Details concerning where the two datasets used for this research can be found as follows: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset and https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis.

Acknowledgments

We acknowledge the administrative and technical support given by Mr. Shehu Abdullahi and the head of department Prof. Idris Ismaila during the period of this research.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

Artificial intelligence assisted in paraphrasing some part of the document.

References

1. Balakrishnan, V.; Khan, S.; Arabnia, H.R. Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Comput. Secur.* **2020**, *90*, 101710.
2. Siddhartha, K.; Raj Kumar, K.; Jayanth Varma, K.; et al. Cyber Bullying Detection Using Machine Learning. In Proceedings of the 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 26–28 August 2022; pp. 1–4.
3. Cuzcano, X.M.; Ayma, V.H. A comparison of classification models to detect cyberbullying in the Peruvian Spanish language on Twitter. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 12–78.
4. Aggarwal, P.; Mahajan, R. Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification. *J. Inf. Syst. Inform.* **2022**, *6*, 607–623.
5. Teng, T.H.; Varathan, K.D. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access* **2022**, *11*, 55533–55560.
6. Vishwamitra, N.; Hu, H.; Luo, F.; et al. Towards understanding and detecting cyberbullying in real-world im-

ages. In Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2021, Online, 21–25 February 2021.
7. Philipo, A.G.; Sarwatt, D.S.; Ding, J.; et al. Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms. *arXiv preprint* **2021**, *arXiv:2407.12154*.
8. Dong, X.; Choi, J.D. XD at SemEval-2020 Task 12: Ensemble approach to offensive language identification in social media using transformer encoders. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 12 December 2020.
9. Dewani, A.; Memon, M.A.; Bhatti, S. Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data. *J. Big Data* **2021**, *8*, 160.
10. Mahesh, K.; Gothane, S.; Toshniwal, A.; et al. Cyber bullying detection on social media using machine learning. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2021**, *7*, 410–416.
11. Dadvar, M.; Trieschnigg, D.; Ordelman, R.; et al. Improving cyberbullying detection with user context. Available online: https://link.springer.com/chapter/10.1007/978-3-642-36973-5_62 (accessed on 23 February 2025).
12. Cirillo, S.; Desiato, D.; Polese, G.; et al. Exploring the ability of emerging large language models to detect cyberbullying in social posts through new prompt-based classification approaches. *Inf. Process. Manag.* **2025**, *62*, 104043. [CrossRef]
13. Agbaje, M.; Afolabi, O. Neural network-based cyber-bullying and cyber-aggression detection using Twitter(X) text. *Rev. Intell. Artif.* **2024**, *38*, 837–846. [CrossRef]
14. Alqahtani, A.F.; Ilyas, M. A machine learning ensemble model for the detection of cyberbullying. *arXiv preprint* **2024**, *arXiv:2402.12538*. [CrossRef]
15. Philipo, A.G.; Sarwatt, D.S.; Ding, J.; et al. Assessing text classification methods for cyberbullying detection on social media platforms. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 7602–7616. [CrossRef]
16. Brakas, K.J.; Alanezi, M. Measuring the extent of cyberbullying comments in Facebook groups for Mosul University students. *Mesopotamian J. Cybersecur.* **2025**, *5*, 337–348.
17. Reynolds, K.; Kontostathis, A.; Edwards, L. Using machine learning to detect cyberbullying. In Proceedings of the 2011 10th International Conference on Machine Learning and Applications, Honolulu, HI, USA, 18–21 December 2011; pp. 241–244.
18. Jadhav, P.A.; Nakhate, M.; Bulani, R.; et al. Real time cyberbullying detection using ML and NLP. *Int. J. Creat. Res. Thoughts* **2023**, *11*, a545–a549.
19. Atoum, J.O. Cyberbullying detection through sentiment analysis. In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence, Las Vegas, NV, USA, 16–18 December 2020; pp. 292–297. [CrossRef]
20. Bharadwaj, V.Y.; Likhitha, V.; Vardhini, V.; et al. Automated cyberbullying activity detection using machine learning algorithm. *E3S Web Conf.* **2023**, *430*, 01039. [CrossRef]
21. Vijayakumar, V.; Prasad, H.D.; Adolf, P. Multimodal cyberbullying detection using hybrid deep learning algorithms. *Int. J. Appl. Eng. Res.* **2021**, *16*, 568.
22. Al-garadi, M.; Hussain, M.R.; Khan, N.; et al. Prediction of cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. *IEEE Access* **2019**, *7*, 70701–70718.
23. Gupta, S.; Vadgama, U.; Vedhavathy, T.R.; et al. Identification and labeling of textual cyberbullying using BiLSTM and BERT. In Proceedings of the 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 5–6 April 2023; pp. 1–5.
24. Shi, L.; Liu, X.; Xu, C.; et al. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Comput. Electr. Eng.* **2021**, *101*, 108005.
25. Akhter, A.; Acharjee, U.K.; Talukder, M.A.; et al. A robust hybrid machine learning model for Bengali cyber bullying detection in social media. *Nat. Lang. Process.* **2023**, *4*, 100027.
26. Al-Hashedi, M.; Soon, L.K.; Goh, H.N. Cyberbullying detection using deep learning and word embeddings: An empirical study. In Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, Bangkok, Thailand, 23–25 November 2019; pp. 17–21.
27. Ali, A.; Syed, A.M. Cyberbullying detection using machine learning. *Pak. J. Eng. Technol.* **2022**, *3*, 45–50. [CrossRef]
28. Hande, A.; Hegde, S.U.; Priyadharshini, R.; et al. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages. *arXiv preprint* **2021**, *arXiv:2108.03867*.

29. Islam, M.M.; Uddin, M.A.; Islam, L.; et al. Cyberbullying detection on social networks using machine learning approaches. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020. [CrossRef]
30. Kumar, A.; Sachdeva, N. Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. *Multimed. Syst.* **2022**, *28*, 2043–2052.
31. Muneer, A.; Alwadain, A.; Ragab, M.G.; et al. Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information* **2023**, *14*, 467.
32. Raj, M.; Singh, S.; Solanki, K.; et al. An application to detect cyberbullying using machine learning and deep learning techniques. *SN Comput. Sci.* **2022**, *3*, 401.
33. Keni, A.; Deepa, M.; Kini, K.V.; et al. Cyber-bullying detection using machine learning algorithms. *Int. J. Eng. Res. Technol.* **2020**, *11*, 690–695.
34. Hani, J.; Nashaat, M.; Ahmed, M.; et al. Social media cyberbullying detection using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 83–87.
35. Thorat, P.B.; Kangane, S.V.; Indalkar, S.V.; et al. Detection of cyber bullying on social media using machine learning. *Int. J. Nov. Res. Dev.* **2022**, *7*, 34–98.
36. Mehendale, N.; Rajpara, K.; Shah, K.; et al. A review on cyberbullying detection using machine learning. Available online: https://ssrn.com/abstract=4116153 (accessed on 13 September 2025).
37. Zhong, J.; Qiu, J.; Sun, M.; et al. To be ethical and responsible digital citizens or not: A linguistic analysis of cyberbullying on social media. *Front. Psychol.* **2021**, *13*, 861823.
38. Nahar, K.M.O.; Alauthman, M.; Yonbawi, S.; et al. Cyberbullying Detection and Recognition with Type Determination Based on Machine Learning. *Comput. Mater. Contin.* **2021**, *75*, 5307–5319.
39. El Koshiry, A.M.; Eliwa, E.H.I.; El-Hafeez, T.A.; et al. Detecting cyberbullying using deep learning techniques: A pre-trained glove and focal loss technique. *PeerJ Comput. Sci.* **2024**, *10*, e1961.