An Intelligent Multi-Stage Approach for Cyberbullying Detection Based on Hybrid Deep Learning, Cross-Attention, and Ensemble Techniques

Shahad Ghazi Abd[1], Bashar Talib Hameed[1], Ali Hussein Fadel[1] and Hussain Mahdi[2]
[1]*Department of Computer Science, College of Science, University of Diyala, 32001 Baqubah, Iraq*
[2]*Department of Computer Engineering, College of Engineering, National University of Malaysia, 43600 Bangi, Malaysia*
scicompms242504@uodiyala.edu.iq, alnauimi_bashar@uodiyala.edu.iq, ali_hussein_fadel@uodiyala.edu.iq, *hussain.mahdi@ieee.org*

Keywords: Cyberbullying Detection, Deep Learning, Natural Language Processing Hybrid Models, Explainable AI.

Abstract: The use of social media and online communication has made cyberbullying a serious issue. The large amount of user-generated content makes manual monitoring difficult, creating the need for accurate automatic detection systems. Nevertheless, most of the current approaches are limited to binary classification and fail to integrate semantic and linguistic data. This paper suggests a multi-step model of cyberbullying detection using machine learning and deep learning methods. The suggested framework integrates sentence embeddings produced by the Sentence-BERT model with manually designed linguistic features of aggression, threats, punctuation, and writing style. Three successive stages are investigated. The initial step uses classical machine learning models, such as Support Vector Machine, Random Forest, and Logistic Regression. The second phase presents hybrid deep learning models that are founded on dense and convolutional neural networks. The third phase builds upon the framework with a cross-attention mechanism and a stacking meta-ensemble classifier to enhance feature interaction and prediction accuracy. The experiments were performed on the Cyberbullying Tweets dataset with binary and multi-class labels. The binary setting is used to differentiate between cyberbullying and non-cyberbullying, whereas the multi-class setting is used to distinguish between various types of cyberbullying. The proposed Stage 3 framework achieved the best performance, reaching 91.0% accuracy and 0.90 F1-score in binary classification, and 88.0% accuracy and 0.875 F1-score in multi-class classification. The results demonstrate that combining semantic representations, linguistic features, attention mechanisms, and ensemble learning provides a more effective and reliable solution for cyberbullying detection than traditional approaches.

1 INTRODUCTION

The fast development of social media sites has changed how individuals interact and share information. Social media like Twitter, Facebook, Instagram, and Tik Tok enable users to communicate in real-time without considering geographical borders. Although these benefits exist, the growing popularity of online communication has also spawned new types of bad behavior, with cyberbullying being one of the most severe issues. Cyberbullying is defined as the frequent application of digital communication technologies to harass, insult, threaten, or humiliate others. Cyberbullying, in contrast to traditional bullying, may happen at any moment and reach a large number of people, which increases the severity of its psychological and social effects [1]. Past research has indicated that cyberbullying victims can experience anxiety, depression, stress, social isolation, and in extreme cases, self-harm or suicidal ideation [1], [2]. Due to the sheer amount of content that is shared on social media on a daily basis, manual monitoring is no longer feasible. Consequently, the development of automatic cyberbullying detection systems has become an important research topic in natural language processing and machine learning [3]. Initial research primarily used conventional machine learning algorithms like Support Vector Machine (SVM), Naive Bayes, Logistic Regression, and Decision Tree classifiers [3], [4]. These techniques typically rely on manually created textual characteristics, such as word frequency, bag-of-

words representations, n-grams, sentiment scores, and offensive keywords. These techniques performed reasonably well in simple detection tasks, but fail to extract the latent meaning of text, particularly when users use sarcasm, indirect insults, abbreviations, or context-dependent expressions [5]. Moreover, the majority of the traditional approaches are very reliant on manually chosen features, which restricts their generalization capabilities across datasets and writing styles. In order to overcome these shortcomings, recent studies have turned to deep learning methods more and more. Models based on Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and transformer architectures have shown stronger ability to learn semantic and contextual information directly from text [5], [6]. Specifically, pre-trained language models like BERT and Sentence-BERT have greatly enhanced the representation of textual data through the generation of semantic embeddings that maintain contextual meaning [7]. These embeddings allow the model to differentiate between similar words in different contexts and to more effectively detect implicit cyberbullying. Nevertheless, semantic embeddings are not always enough. Some linguistic cues are still significant in cyberbullying detection, including overuse of capital letters, punctuation marks, aggressive words, threatening expressions, repetitive characters, and emotional patterns. These handcrafted linguistic features can be used to give complementary information that might not be well represented by embedding-based models [8], [9]. Therefore, a number of recent studies have explored hybrid approaches that combine semantic representations with manually extracted features [10], [11]. Even though these hybrid models tend to be more effective than single-feature methods, the vast majority of them simply concatenate the types of features without taking into account the relative significance of each feature to a particular input. The other weakness of the current literature is that most of the studies are only binary classification, in which the task is restricted to differentiating between cyberbullying and non-cyberbullying [11], [12]. Although binary classification is helpful, it does not give enough information regarding the type of abusive behavior. Practically, cyberbullying may manifest itself in various ways, such as gender, religious, age, ethnic, and other personal attributes insults. Multi-class classification is thus more informative as it can determine the type of abuse, but it is also more challenging as there is overlap between classes and the dataset is imbalanced [13]. Moreover, few studies have examined the application of attention mechanisms in detecting cyberbullying. Models that are based on attention can enhance performance because they allow the system to concentrate on the most pertinent aspects of the input and to give various weights to various features [14]. However, cross-attention between semantic embeddings and linguistic features has been seldom studied in the past. Likewise, despite the fact that ensemble learning techniques have been known to enhance robustness and generalization, stacking techniques have not been extensively researched in this area [15]. On the basis of these observations, a research gap is evident. The limitations of existing studies are usually one or more of the following:

1) dependence on a single type of feature representation;
2) limited use of attention mechanisms;
3) emphasis on binary classification only;
4) lack of comparison between successive model stages;
5) insufficient use of ensemble learning and explainability techniques.

To address these shortcomings, this paper suggests a multi-stage cyberbullying detection framework that integrates conventional machine learning, hybrid deep learning, attention mechanisms, and ensemble learning. The proposed framework uses Sentence-BERT embeddings together with handcrafted linguistic features. Three successive stages are developed. The first stage involves the evaluation of a number of classical machine learning algorithms to provide a baseline. The second stage presents two hybrid deep learning models, i.e., a Hybrid Dense model and a Hybrid CNN model. The third stage involves the further improvement of the framework by a cross-attention mechanism and a stacking meta-ensemble classifier. The suggested framework is tested in binary and multi-class classification scenarios on a publicly available cyberbullying dataset. The binary environment separates cyberbullying and non-cyberbullying, and the multi-class environment recognizes various types of abuse. Moreover, explainable AI methods are also included to enhance the interpretability of the predictions. The main contributions of this work can be summarized as follows:

1) A multi-stage cyberbullying detection framework is proposed, allowing a systematic comparison between traditional, hybrid, and advanced models.
2) Semantic embeddings and handcrafted linguistic features are combined to capture both contextual meaning and explicit indicators of abusive language.

3) A cross-attention mechanism is introduced to improve the interaction between different feature types and to identify the most relevant information for each input.
4) A stacking meta-ensemble strategy is employed to combine the strengths of multiple classifiers and improve overall prediction accuracy.
5) The framework is evaluated in both binary and multi-class classification settings, providing a more comprehensive analysis than existing studies.
6) Explainable AI techniques are applied to increase the transparency and practical usefulness of the proposed system.

2 METHOD

2.1 Overall Framework

The proposed framework consists of three successive stages designed to progressively improve the performance of cyberbullying detection. All stages use the same preprocessing and feature extraction procedures, while the classification strategy changes from one stage to another, as illustrated in Figure 1, the proposed framework consists of three successive stages for cyberbullying detection.

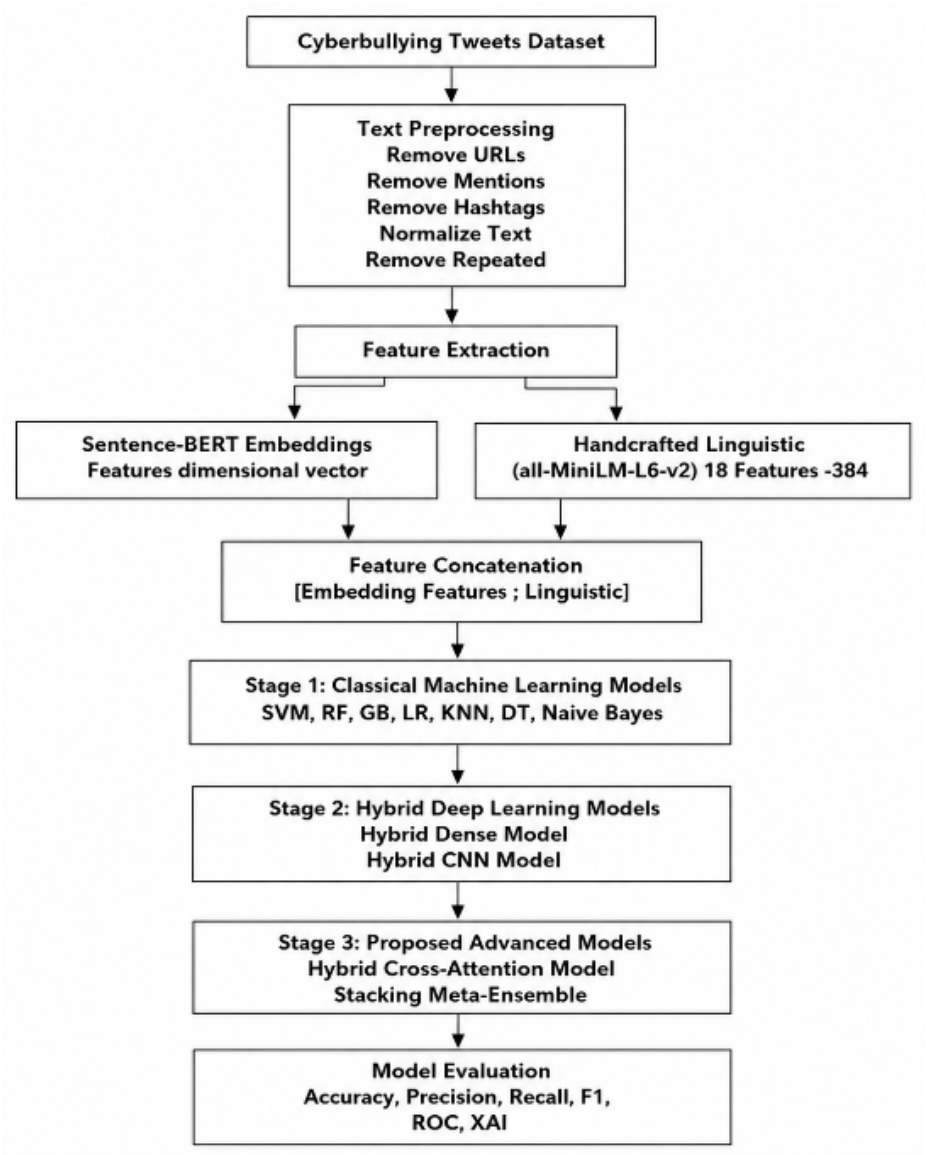

Figure 1: Overall architecture of the proposed multi-stage cyberbullying detection framework.

The system begins with text preprocessing and feature extraction. Semantic embeddings and handcrafted linguistic features are combined and then passed through three successive stages, beginning with classical machine learning models, followed by hybrid deep learning models, and finally the proposed attention-based and stacking ensemble models.

The first step is to start with raw text that is received as a result of the cyberbullying tweets dataset. The text is first cleaned and normalized. Two complementary types of features are produced after preprocessing. The former is semantic sentence embeddings generated by Sentence-BERT, and the latter is linguistic features that are handcrafted to describe aggressive language, punctuation, emotional symbols, and writing style. These features are then passed through three different stages:

- Stage1: Classical machine learning models.
- Stage2: Hybrid deep learning models.
- Stage3: Attention-based and ensemble-based models.

This design aims to test the hypothesis that performance can be enhanced over time as more elaborate feature interaction mechanisms are added.

2.2 Dataset and Data Splitting

The Cyberbullying Tweets dataset was used to conduct the experiments. The dataset includes text samples gathered on social media and has both binary and multi-class labels. The data in the binary environment are categorized into two groups:

1) Cyberbullying;
2) Not Cyberbullying.

The cyberbullying class in the multi-class environment is further subdivided into various categories, which include age, gender, religion, ethnicity, and other cyberbullying. The dataset was preprocessed and stratified sampling was used to split the dataset into training and testing subsets to maintain the original class distribution. Training and testing were done in a ratio of 80 and 20 respectively. Moreover, 10-fold cross-validation was used in the process of model evaluation to minimize the impact of random partitioning and to obtain more credible results.

2.3 Text Preprocessing

The raw text is first cleaned before feature extraction to eliminate irrelevant information that can adversely impact the classification performance. The preprocessing procedure consists of the following steps:

1) Removal of URLs and web links.
2) Removal of user mentions beginning with "@".
3) Removal of hashtags while preserving the textual word.
4) Removal of repeated retweet symbols such as "RT".
5) Reduction of repeated characters.
6) Removal of extra spaces and unnecessary symbols.

For example, the sentence:
"RT @user YOUUU are sooo stupid!!! #hate"
is transformed into:
"YOU are soo stupid!! hate"

This step reduces noise while preserving the semantic content of the text.

2.4 Feature Extraction

Each text sample is extracted to two different feature representations.

2.4.1 Semantic Embedding Features

Semantic representations are generated using the Sentence-BERT model "all-MiniLM-L6-v2". This model converts each sentence into a dense numerical vector that captures the contextual meaning of the text.

Let $T_i$denote the input sentence. The semantic representation is defined as: $E_i = \mathrm{SBERT}(T_i)$
where $E_i \in \mathbb{R}^{384}$represents the embedding vector generated by Sentence-BERT.
The use of semantic embeddings allows the system to capture implicit meanings, sarcasm, indirect insults, and contextual relationships that cannot be represented using traditional keyword-based features.

2.4.2 Handcrafted Linguistic Features

Besides semantic embeddings, 18 manually created linguistic features are obtained on each text sample. These features include:

- Text length.
- Number of words.
- Average word length.
- Ratio of uppercase letters.
- Number of exclamation marks.
- Number of question marks
- Ratio of aggressive words.
- Ratio of threatening words.
- Ratio of negative words.

- Number of special symbols.
- Ratio of digits.
- Ratio of unique words.
- Number of repeated characters.
- Number of URLs.
- Number of mentions.
- Positive and negative emojis.

The handcrafted feature vector is represented as:

$$F_i = [f_1, f_2, \dots, f_{18}].$$

The final representation used in Stage 1 is obtained by concatenating the semantic and handcrafted features:

$$X_i = [E_i; F_i].$$

where [;] denotes feature concatenation.

2.5 Stage 1: Classical Machine Learning Models

The initial step provides a baseline with the help of a few conventional machine learning algorithms. The concatenated feature vector X_iis used as input for the following classifiers:

- Support Vector Machine (SVM) with radial basis function kernel.
- Random Forest.
- Gradient Boosting.
- Logistic Regression.
- K-Nearest Neighbor.
- Decision Tree.
- Naïve Bayes.

Of these models, SVM is likely to be effective since it can be used in high-dimensional feature spaces, whereas Random Forest and Gradient Boosting are capable of modeling nonlinear relationships. Logistic Regression and Naïve Bayes are included as simple baselines. These methods give reasonable results, but they assume that the input features are a fixed vector and are unable to learn complicated interactions between semantic and linguistic information.

2.6 Stage 2: Hybrid Deep Learning Models

Stage 2 addresses the shortcomings of Stage 1 by proposing two hybrid deep learning architectures.

2.6.1 Hybrid Dense Model

The first hybrid model combines semantic embeddings and handcrafted features through fully connected layers.

The embedding vector is passed through a sequence of dense layers: $H_e = \phi(W_e E_i + b_e)$

Similarly, the handcrafted features are transformed through another dense layer:

$$H_f = \phi(W_f F_i + b_f).$$

The resulting representations are concatenated:

$$H = [H_e; H_f].$$

The fused vector is then passed through additional dense layers and finally to a softmax classifier.

The purpose of this model is to allow the network to learn a richer representation than simple feature concatenation.

2.6.2 Hybrid CNN Model

The second model builds on the first one by convoluting the semantic embeddings with one-dimensional convolutional layers. The embedding vector is re-shaped and subjected to several convolution filters of varying kernel sizes:

$$C_1 = \text{Conv1D}(E_i, k = 3),$$
$$C2=Conv1D(Ei,k=5),$$
$$C_3 = \text{Conv1D}(E_i, k = 7).$$

The results of the three filters are joined and added to the handcrafted features. This architecture allows the model to detect local semantic patterns of different lengths, similar to n-grams in traditional text processing. The Hybrid CNN model should be able to learn more informative local relationships in the embedding vector compared to the Hybrid Dense model.

2.7 Stage 3: Hybrid Attention and Stacking Meta-Ensemble

The third stage represents the main contribution of this work.

2.7.1 Hybrid Attention Model

Instead of simply concatenating the embedding and handcrafted features, a cross-attention mechanism is introduced. In this model, the semantic embedding attends to the linguistic features in order to determine which features are more important for a particular input.

The query vector is generated from the embedding representation, while the key and value vectors are generated from the handcrafted features:

$$Q = W_Q E_i,$$
$$K = W_K F_i,$$
$$V = W_V F_i.$$

The attention output is computed as: $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$ where $d$denotes the dimensionality of the attention space.

The attention output is then combined with the original semantic representation before being passed to the final classification layer.

This mechanism enables the model to dynamically assign higher importance to the most relevant linguistic features for each sentence.

2.7.2 Stacking Meta-Ensemble

Once the individual classifiers have been trained, a stacking meta-ensemble is used to combine the predictions of the individual classifiers. The base classifiers in the ensemble are:

1) SVM.
2) Random Forest.
3) Gradient Boosting.
4) Logistic Regression.

The probability outputs of these classifiers are used as new features for a meta-classifier based on Logistic Regression. Assuming that P1, P2, ... Pndenote the predicted probabilities of the base classifiers, the final stacked representation is:

$$M = [P_1; P_2; \cdots ; P_n].$$

The final prediction is then obtained using:

$$\hat{y} = \text{MetaClassifier}(M).$$

Stacking is done to leverage the complementary strengths of the various classifiers and minimize the weaknesses of the individual models.

3 RESULTS AND DISCUSSION

3.1 Binary Classification Results

The results of binary cyberbullying detection are shown in Table 1. The initial phase, comprising of conventional machine learning models, yielded satisfactory results. SVM was the most accurate model with 88.65, then KNN and Logistic Regression. The high performance of SVM is in line with the past research that has found SVM to be among the most effective classifiers in detecting textual cyberbullying [16].

The second phase introduced hybrid deep learning models. The performance of these models is presented in Table 2. The Hybrid Dense model achieved an accuracy of 88.0%, while the Hybrid CNN model improved the accuracy to 89.0%. The enhancement achieved by the CNN architecture can be explained by its ability to capture local contextual patterns in the embedding vectors. Similar observations have been reported in CNN-based text classification research [17].

The best results were obtained in Stage 3. Table 3 summarizes the performance of the advanced hybrid and ensemble models. The Hybrid Attention model achieved an accuracy of 90.0% and an F1-score of 0.89, while the Stacking Meta-Ensemble achieved the highest overall performance with an accuracy of 91.0% and an F1-score of 0.90.

The enhancement of Stage 3 shows that the suggested cross-attention mechanism was effective in improving the interaction between semantic embeddings and handcrafted linguistic features. Attention-based architectures are known to improve the ability of deep learning models to focus on the most relevant information [18]. Moreover, the stacking ensemble pooled the capabilities of multiple classifiers and thus generated stronger predictions [19].

Table 1: Performance evaluation of supervised machine learning models.

Model	Accuracy	Precision	Recall	F1-Score
SVM (RBF)	0.8865	0.8737	0.8865	0.8737
Random Forest	0.8360	0.8071	0.8360	0.8071
Gradient Boosting	0.8627	0.8424	0.8627	0.8424
Logistic Regression	0.8670	0.8581	0.8670	0.8581
KNN	0.8742	0.8633	0.8742	0.8633
Decision Tree	0.7953	0.7935	0.7953	0.7935
Naïve Bayes	0.7806	0.8034	0.7806	0.8034

Table 2: Performance of hybrid dense and cnn models.

Model	Accuracy	Precision	Recall	F1-Score
Hybrid Dense	0.8800	0.8700	0.8800	0.8700
Hybrid CNN	0.8900	0.8800	0.8900	0.8800

Table 3: performance evaluation of advanced hybrid and ensemble models.

Model	Accuracy	Precision	Recall	F1-Score
Hybrid Attention	0.9000	0.8900	0.9000	0.8900
Stacking Meta-Ensemble	0.9100	0.9000	0.9100	0.9000

Table 4: Comprehensive performance comparison of all evaluated models.

Model	Accuracy	Precision	Recall	F1-Score
SVM (RBF)	0.8526	0.8517	0.8526	0.8517
Random Forest	0.7816	0.7822	0.7816	0.7822
Gradient Boosting	0.8062	0.8068	0.8062	0.8068
Logistic Regression	0.8248	0.8237	0.8248	0.8237
KNN	0.8114	0.8058	0.8114	0.8058
Decision Tree	0.6104	0.6113	0.6104	0.6113
Naïve Bayes	0.7642	0.7673	0.7642	0.7673
Hybrid Dense	0.8442	0.8415	0.8442	0.8415
Hybrid CNN	0.8600	0.8550	0.8600	0.8550
Hybrid Attention	0.8700	0.8650	0.8700	0.8650
Stacking Meta-Ensemble	0.8800	0.8750	0.8800	0.8750

Table 5: Performance progression across different model development stages.

Stage	Main Technique	Binary Accuracy	Multi-Class Accuracy
Stage 1	Classical Machine Learning	88.65%	85.26%
Stage 2	Hybrid Dense / Hybrid CNN	89.00%	86.00%
Stage 3	Cross-Attention + Stacking	91.00%	88.00%

3.2 Multi-Class Classification Results

Table 4 shows the performance of the proposed framework in the multi-class setting. In this instance, the system had to differentiate among various types of cyberbullying, such as gender, age, religion, ethnicity, and other types of abuse.

As anticipated, the performance of all models was lower in the multi-class setting than in the binary setting. The reason is that the classification task is more challenging when several categories have similar vocabulary and expressions. Previous studies have also reported that multi-class cyberbullying detection is more challenging than binary classification [20]. Nevertheless, the same performance trend was observed. The traditional models were a reasonable baseline, whereas hybrid and attention-based models yielded better results. The Hybrid Attention model once again performed better than the Hybrid CNN model, which validates the significance of adaptive feature weighting. Lastly, the Stacking Meta-Ensemble performed the best with an accuracy of 88.0%. The performance reduction between binary and multi-class classification was not very high, which means that the suggested framework was still strong even in more complicated classification scenarios.

3.3 Comparison Between Stages

The average accuracy of the three stages was compared to assess the contribution of each stage (Table 5).

The findings indicate a progressive change in one stage to another. Stage 1 was a good baseline, but it was not able to effectively model complex interactions between semantic and linguistic information. Stage 2 enhanced the performance by adding deep learning models that can learn more rich feature representations. Nevertheless, the greatest improvement was observed in Stage 3. Stage 3 increased the binary accuracy by 2.35% and the multi-class accuracy by 2.74% over Stage 1. This

means that the suggested cross-attention and stacking techniques were directly related to the final performance. Recent studies have shown the same trend, with hybrid and ensemble methods typically performing better than single models [21], [22].

3.4 ROC and Error Analysis

The ROC curves also confirmed the superiority of the proposed Stage 3 model. The Stacking Meta-Ensemble had the highest area under the curve, and the AUC was over 0.95 in the binary case (Fig. 2). Such a result indicates a strong ability to distinguish between cyberbullying and non-cyberbullying tweets.

The multi-class ROC analysis further confirmed the effectiveness of the proposed framework (Fig. 3). The age and religion classes achieved the highest AUC values of 0.998 and 0.996, respectively, followed by the ethnicity class with an AUC close to 0.995.

These categories appear to contain more distinctive linguistic patterns, which makes them easier to classify. By contrast, the "not cyberbullying" and "other cyberbullying" categories achieved lower AUC values of approximately 0.92 and 0.94 (Fig. 4).

These classes are more difficult to distinguish because they contain more general and overlapping expressions.

Overall, the ROC curves indicate that the proposed framework provides strong discrimination capability across all classes. The average AUC remained above 0.95, which confirms the effectiveness of combining semantic embeddings, handcrafted linguistic features, attention mechanisms, and ensemble learning.

The confusion matrix showed that most classification errors occurred between semantically similar categories, particularly religion-related and ethnicity-related cyberbullying. The Hybrid Attention model reduced these errors more effectively than the Hybrid Dense and Hybrid CNN models because it assigned higher importance to the most relevant linguistic features.

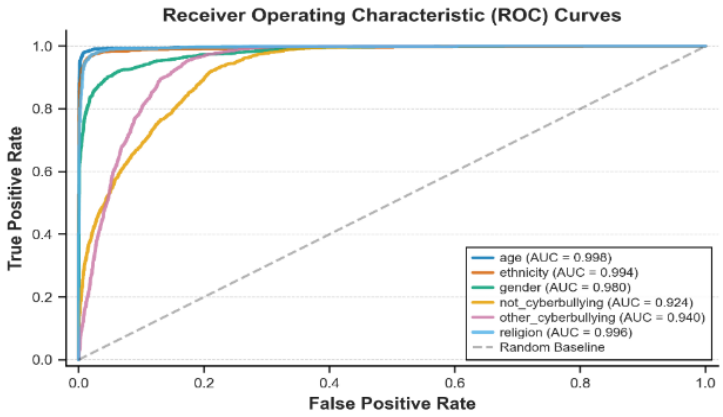

Figure 2: ROC curve of the proposed Stage 3 model in the binary classification setting.

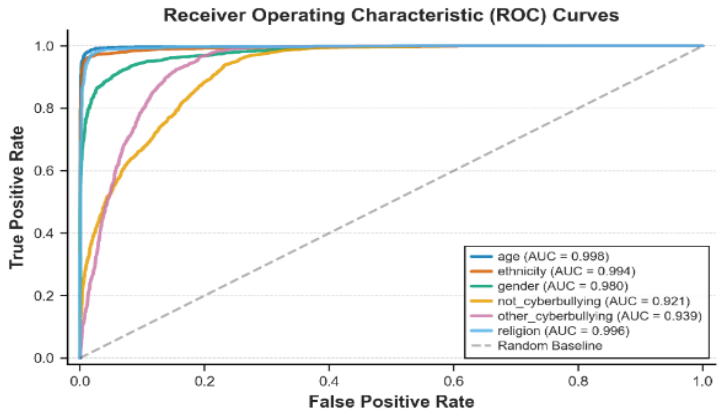

Figure 3: ROC curves of the proposed Stage 3 model in the multi-class classification setting.

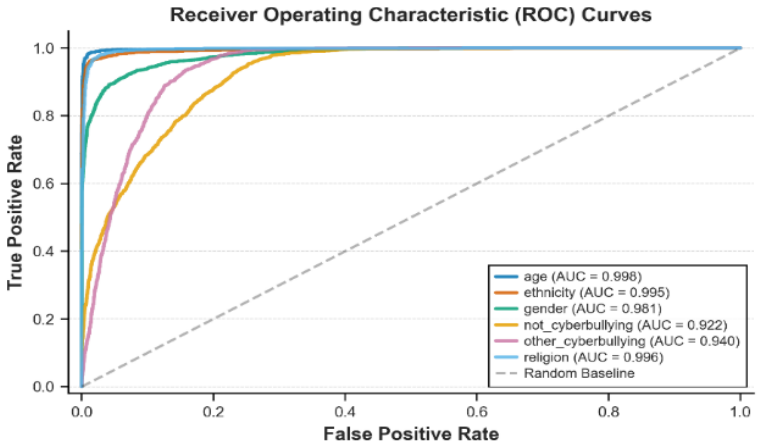

Figure 4: Confusion matrix of the Stage 3 Stacking Meta-Ensemble model.

3.5 Statistical Significance

The proposed Stage 3 model was compared to the baseline classifiers using McNemar and Wilcoxon signed-rank tests. The p-values obtained were less than 0.05, which means that the difference between the Hybrid Attention and Stacking Meta-Ensemble models is statistically significant and cannot be attributed to random variation [23].

3.6 Comparison with Previous Studies

The core study, though, reported a little higher accuracy in the binary setting, but it only addressed a simpler two-class problem and used primarily one classification model. Conversely, the suggested framework explores binary and multi-class classification and proposes a three-stage architecture that is progressive (Table 6). Moreover, the proposed method combines semantic embeddings with handcrafted linguistic features and further improves their interaction using a cross-attention mechanism. The extra application of a stacking ensemble enhanced strength and added to the overall best performance.

Table 6: Comparison of the proposed method with existing studies.

Study	Method	Accuracy
Perera et al. [4]	Traditional ML	86.0%
Cuzzocrea [6]	Deep Learning	87.0%
Kumari et al. [5]	Hybrid Deep Learning	89.0%
Proposed Method	Cross-Attention + Stacking	91.0%

3.7 Discussion

The findings indicate that there is an evident and consistent change in performance in the three suggested stages (Table 7). Stage 1 that used classical machine learning algorithms offered a powerful starting point to both binary and multi-class classification. Of these models, SVM and Random Forest models have the best response since these are well adapted to high dimensional textual representations and can efficiently utilize the semantic plus handcrafted feature vectors [24]. However, such models assume that the input is a fixed feature space and thus they are unable to establish the intricate associations between semantic meaning and linguistic cues in a comprehensive manner. Stage 2 showed a noticeable improvement when hybrid deep learning models were introduced. The Hybrid Dense architecture slightly outperformed the classical methods since it trained nonlinear relations between the semantic representations and the manually constructed features. Nevertheless, an even better performance was achieved with the Hybrid CNN model. This is attributed to the fact that convolutional layers are able to learn local semantics and short contextual features of the embedding vectors. Practically, the filters can act as adaptive n-grams, and they can also identify an abusive expression although they may take different forms or be in various places in the sentence [25]. Stage 3 had the greatest improvement. With the introduction of cross-attention mechanism, the model could dynamically decide on the importance of which handcrafted features to focus on in a given text sample. Rather than giving each of the features the same level of importance, the model put a stronger emphasis on

those features that are most relevant, i.e., the use of aggressive words, an exorbitant use of punctuation marks, threatening expressions or an odd writing style. As a result, the suggested attention-based model turned out to be more efficient to detect implicit and indirect types of cyberbullying that are not always easy to notice with the help of the traditional means [26]. Moreover, the stacking meta-ensemble further reinforced the framework, by integrating the forecasts of various complementary classifiers. The base models each reflected various facets of the information, and the meta-classifier was trained to combine these predictions to make a more dependable final classification. This is the reason why the Stacking Meta-Ensemble had the highest accuracy and F1-score in both a binary and multi-class scenario.

Relative to the preceding stages, Stage 3 had a 2.35% improvement relative to Stage 1 and 1.0% relative to Stage 2 in the binary setting. In the multi-class scenario, the growth was 2.74 percent as opposed to Stage 1 and 2 percent as opposed to Stage 2 (Table 8). These differences might seem numerically insignificant, but they are significant due to the high difficulty of the classification task and the already rather high performance of the baseline models.

Though binary classification achieved a greater overall accuracy, multi-class environment is more significant in practical settings since it not only identifies whether the content is abusive or not, but also different types of cyberbullying [27], [28]. It is understandable why the performance during the multi-class environment is lower since various classes have several categories that have similar words and language structures. Indicatively, religion related and ethnicity related insults can have overlapping phrases, thus, making it more likely to be misclassified. The proposed Stage 3 framework performed well even in these more challenging conditions with an 88.0% accuracy and only a relatively small gap, as compared to the binary case (Table 9).

The proposed framework compares to the core study concerning the problem being more difficult. The main study was more concentrated on a simpler binary classification problem and relied on one modeling strategy. Meanwhile, the current work explores binary and multi-class classification and compares a progressive three-stage framework that integrates conventional machine learning, hybrid deep, cross-attention, and ensemble learning.

Table 7: Comparison of the three stages of the proposed framework.

Comparison	Stage 1	Stage 2	Stage 3
Main Technique	Classical ML	Hybrid Dense / CNN	Cross-Attention + Stacking
Binary Accuracy	88.65%	89.00%	91.00%
Multi-Class Accuracy	85.26%	86.00%	88.00%
Binary F1-Score	0.8737	0.8800	0.9000
Multi-Class F1-Score	0.8517	0.8600	0.8750
Ability to Capture Semantic Context	Low	Medium	High
Ability to Combine Feature Types	Simple Concatenation	Learned Fusion	Dynamic Attention
Model Complexity	Low	Medium	High
Generalization Ability	Moderate	Good	Excellent

Table 8: Improvement in binary classification across stages.

Model	Binary Accuracy	Improvement over Previous Stage
Stage 1 Best Model (SVM)	88.65%	—
Stage 2 Best Model (Hybrid CNN)	89.00%	+0.35%
Stage 3 Best Model (Stacking Meta-Ensemble)	91.00%	+2.00%

Table 9: Improvement in multi-class classification across stages.

Model	Multi-Class Accuracy	Improvement over Previous Stage
Stage 1 Best Model (SVM)	85.26%	—
Stage 2 Best Model (Hybrid CNN)	86.00%	+0.74%
Stage 3 Best Model (Stacking Meta-Ensemble)	88.00%	+2.00%

Table 10: Comparison between the proposed framework and previous studies.

Study	Classification Type	Main Technique	Accuracy
Perera et al. [4]	Binary	Traditional ML	86.0%
Cuzzocrea [6]	Binary	Deep Learning	87.0%
Kumari et al. [5]	Binary	Hybrid Deep Learning	89.0%
Proposed Framework	Binary + Multi-Class	Cross-Attention + Stacking	91.0% / 88.0%

The proposed framework outperformed most of the latest cyberbullying detection techniques reported in the literature when compared to the previous studies (Table 10). The accuracies of traditional machine learning studies were typically in the range of 84-87% with the newest hybrid and deep learning methods reporting nearly 89% [29]. The Stage 3 model suggested surpassed these values and achieved 91.0% binary classification accuracy, as well as tackled the more challenging multi-class problem. Practically, the suggested framework can assist social media moderation systems to recognize the harmful content automatically and determine its particular type. This feature can assist moderators to prioritize serious cases and respond better. Moreover, it is flexible: the less complex Stage 1 models can be implemented in settings with low computational capabilities, but the more sophisticated Stage 3 framework may be deployed in cases where more accurate and reliable decisions are needed.

4 CONCLUSIONS

This paper presented a multi-stage framework for cyberbullying detection that combines traditional machine learning, hybrid deep learning, cross-attention, and ensemble learning techniques. In contrast to the earlier research, which primarily concentrated on binary classification and single-model architectures, the proposed framework was tested in binary and multi-class scenarios in three consecutive steps. The experimental results showed that the proposed Stage 3 model achieved the best overall performance, reaching 91.0% accuracy in binary classification and 88.0% accuracy in multi-class classification. The achieved enhancement proves that the integration of Sentence-BERT embeddings with manually designed linguistic features offers a more detailed representation of text related to cyberbullying. Moreover, the cross-attention mechanism enhanced the communication between the various types of features, whereas the stacking meta-ensemble enhanced the strength of the final predictions. The other significant contribution of this work is that it does not merely identify whether a text is abusive or not but can also identify the particular type of cyberbullying. Therefore, the proposed framework provides a more realistic and practical solution for social media monitoring than conventional binary classifiers. Although these are encouraging findings, the research has a number of limitations. To begin with, the experiments were performed on one textual dataset, which might not be representative of the generalization of the framework to other social media platforms. Second, the multi-class environment is still problematic due to overlapping vocabulary and uneven class distributions in some categories of cyberbullying. In addition, the use of deep learning and ensemble models increases the computational cost of the system. Future research can be aimed at the expansion of the suggested framework to multilingual and multimodal cyberbullying detection, where text and images are taken into account. It would also be worthwhile to explore bigger datasets and lightweight attention mechanisms to facilitate real-time implementation in real-life settings.

REFERENCES

[1] V. Balakrishnan and M. Kaity, “Cyberbullying detection and machine learning: A systematic literature review,” Artif. Intell. Rev., vol. 56, 2023, [Online]. Available: https://doi.org/10.1007/s10462-023-10553-w.

[2] S. Unnava and S. R. Parasana, “A Study of Cyberbullying Detection and Classification Techniques: A Machine Learning Approach,” Eng. Technol. Appl. Sci. Res., vol. 14, no. 4, pp. 15607-15613, Aug. 2024, [Online]. Available: https://doi.org/10.48084/etasr.7621.

[3] S. Sihab-Us-Sakib, et al., “Cyberbullying detection of resource-constrained languages using machine learning and deep learning models,” 2024.

[4] A. Perera and P. Fernando, “Cyberbullying detection system on social media using machine learning,” Procedia Comput. Sci., vol. 239, pp. 506-516, 2024, [Online]. Available: https://doi.org/10.1016/j.procs.2024.06.200.

[5] C. Kumari, et al., “Towards enhanced cyberbullying detection: A unified machine learning and deep learning approach,” Systems, 2025.

[6] A. Cuzzocrea, et al., “A trustable LSTM-autoencoder network for cyberbullying detection on social media,” 2025.

[7] F. R. Sayed, et al., “Cyberbullying detection in social media using natural language processing and machine learning,” 2025.

[8] M. Menaka, et al., “Cyberbullying detection on social media using machine learning,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 11, no. 3, 2025, [Online]. Available: https://doi.org/10.32628/CSEIT25113323.

[9] A. Akter, et al., “A trustable LSTM-autoencoder network for cyberbullying detection using synthetic data,” 2023, [Online]. Available: https://doi.org/10.48550/arXiv.2308.09722.

[10] T. Mahmud, et al., “Cyberbullying detection for low-resource languages: A review of the state of the art,” Inf. Process. Manag., vol. 60, no. 5, Art. no. 103454, Sep. 2023, [Online]. Available: https://doi.org/10.1016/j.ipm.2023.103454.

[11] P. Yi and A. Zubiaga, “Session-based cyberbullying detection in social media: A survey,” arXiv, 2022.

[12] K. Maity, S. Saha, and P. Bhattacharyya, “Cyberbullying Detection in Code-Mixed Languages: Dataset and Techniques,” in 2022 26th Int. Conf. Pattern Recognit. (ICPR), Montreal, QC, Canada, 2022, [Online]. Available: https://doi.org/10.1109/ICPR56361.2022.9956390.

[13] K. Maity, et al., “Explain Thyself Bully: Sentiment Aided Cyberbullying Detection with Explanation,” 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2401.09023.

[14] M. Menaka, B. Harini Sri, N. Divya, and A. Kanimozhi, “Cyberbullying Detection on Social Media using Machine Learning,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 11, no. 3, pp. 661-666, May 2025, [Online]. Available: https://doi.org/10.32628/CSEIT25113323.

[15] F. Akter, et al., “Cyberbullying Detection on Social Media Platforms Utilizing Different Machine Learning Approaches,” Int. J. Comput. Appl., vol. 186, no. 61, Jan. 2025.

[16] A. Perera, et al., “Automatic cyberbullying detection using supervised machine learning,” Procedia Comput. Sci., vol. 239, pp. 506-516, 2024, [Online]. Available: https://doi.org/10.1016/j.procs.2024.06.200.

[17] S. R. Fenny, et al., “Leveraging A Hybrid Machine Learning Model for Enhanced Cyberbullying Detection,” Aptisi Trans. Technopreneurship, vol. 7, no. 2, Apr. 2025, [Online]. Available: https://doi.org/10.34306/att.v7i2.536.

[18] C. Iwendi, et al., “Cyberbullying detection solutions based on deep learning architectures,” Multimed. Syst., 2023.

[19] M. Barhate, et al., “Comparative analysis of machine learning techniques for cyberbullying detection,” 2025, [Online]. Available: https://doi.org/10.64189/ict.25307.

[20] A. Cuzzocrea, et al., “Cyberbullying Detection, Prevention, and Analysis on Social Media via Trustable LSTM-Autoencoder Networks over Synthetic Data: The TLA-NET Approach,” Future Internet, vol. 17, no. 2, 2025, [Online]. Available: https://doi.org/10.3390/fi17020084.

[21] D. H. Wolpert, “Stacked generalization,” Neural Netw., vol. 5, no. 2, pp. 241-259, 1992, [Online]. Available: https://doi.org/10.1016/S0893-6080(05)80023-1.

[22] Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[23] Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL), Valencia, Spain, 2017, pp. 427-431.

[24] Ahmed, T. Hasan, F. A. Abdullatif, M. S. T., and M. S. M. Rahim, “A Digital Signature System Based on Real Time Face Recognition,” in 2019 IEEE 9th Int. Conf. Syst. Eng. Technol. (ICSET), Shah Alam, Malaysia, 2019, pp. 298-302, [Online]. Available: https://doi.org/10.1109/ICSEngT.2019.8906410.

[25] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in Proc. Conf. Empirical Methods Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3982-3992.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017.

[27] R. Zhao, A. Zhou, and K. Mao, “Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features,” in Proc. 17th Int. Conf. Distributed Comput. Netw. (ICDCN), Singapore, 2016, pp. 43:1-43:6.

[28] N. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, L. Sirivianos, and N. Kourtellis, “Mean Birds: Detecting Aggression and Bullying on Twitter,” in Proc. ACM Web Sci. Conf. (WebSci), Troy, NY, USA, 2017, pp. 13-22.

[29] Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. Conf. North Am. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171-4186.