ISSN: 3079-5354

STAP
Smart Technologies Academic Press

Journal of Cyber Security and Risk Auditing
https://www.jcsra.thestap.com/

Hybrid BERT-XGBoost Framework for Early Detection and Classification of Online Cyberbullying across Social Media

Rima Shishakly[1], Abdelhadi Mohammad Bayoud[2], Mansour Obeidat[3*], Hussein Edrees[4]

[1]*Management & Marketing Department, College of Business Administration, Ajman University, Ajman 346, United Arab Emirates*
[2]*Cybersecurity Consultant, Security Matterz, Riyadh, Saudi Arabia*
[3]*Applied College, King Faisal University, Al-Ahsa, Saudi Arabia*
[4]*Deanship of Development and Quality Assurance, King Faisal University, 31982, Al-Ahsa, Saudi Arabia*

ARTICLE INFO

Article History

Received: 17-02-2026
Revised: 30-04-2026
Accepted: 27-06-2026
Published: 30-06-2026

Vol.2026, No.2

DOI:

***Corresponding author.**
Email:
Mobaydat@kfu.edu.sa

Orcid: https://orcid.org/0009-0002-1649-5573

ABSTRACT

Cyberbullying via social media is a constant digital safety issue because the content can be widely shared and openly visible and can have a negative impact on users before it is removed by manual moderation. Current detection models are mostly based on shallow lexical features or transformer-only classifiers, resulting in low-level accuracy and explainability. This study introduces a Hybrid BERT–XGBoost model to detect cyberbullying in short social media texts, which combines the strengths of both models. The contextual sentence embeddings are extracted using BERT and the auxiliary linguistic and behavioral features are extracted in parallel, such as sentiment polarity, profanity score, punctuation intensity, capitalization ratio, hashtag usage, mention count, emoji frequency, and post length. XGBoost is used for the classification of the fused representation. The model is tested on stratified training, validation, and testing splits, compared to a baseline model, ablated, tested with macro-F1, weighted-F1, ROC-AUC, early detection recall, and grouped explainability. The proposed framework achieved 96.18% accuracy, 96.05% macro-F1, 96.16% weighted-F1, 95.88% early detection recall, and 98.42% macro-AUC. It performs better than the BERT + Dense baseline, which obtained 94.31% accuracy and 94.08% macro-F1 score, demonstrating the advantage of fusion of contextual and auxiliary features. The framework provides an interpretable, practical and category-aware solution for early detection of cyberbullying, but further research is needed to validate the framework in multiple languages, modalities and in conversations.

Keywords: Cyberbullying detection, BERT, XGBoost, Social media text classification, Explainable AI.

How to cite the article

1. Introduction

Online social media has revolutionized the ways in which people communicate, learn, and debate and form communities. The very same openness that is beneficial to these platforms has also allowed for the repeated digital aggression, targeted humiliation, hate-filled comments, identity-based attacks, threats, and coordinated harassment. Cyberbullying is not limited to any one forum, platform, or age group and can be seen on microblogging platforms, messaging platforms, video-sharing platforms, gaming platforms, and comment sections where harmful language can be seen and spread rapidly and for extended periods of time [1]. Online abuse is different to the bullying that happens in schools, workplaces or in the local community because it can be anonymous, seen by many people in seconds and can go on after school, work or local community. Cyberbullying is not just a behavioral issue as it is persistent and can be borderless.

Cyberbullying can have a significant effect since the victim may be subjected to public humiliation, emotional stress, social isolation and multiple psychological trauma. Previous research has found that cyberbullying victimization is associated with anxiety, depression, low self-esteem, social withdrawal and suicidal ideation particularly among adolescents and young adults [2]. The issue gets trickier when negative content is conveyed in slang, sarcasm, emoji, abbreviations, code mixed language, indirect insults or context dependent expressions. A single sentence can seem neutral on its own, but can be abusive when read in context, with the intent of the user, or the vocabulary of the community.

The initial research was predominantly based on lexicon-based rules, bag-of-words features, term frequency–inverse document frequency, sentiment indicators, as well as traditional classifiers like Naïve Bayes, logistic regression, support vector machines, random forests and decision trees [3]. These models are easy to build and are efficient, but they are not always able to understand the deep semantic meaning, implicit aggression, and context differences between platforms. More recent methods used deep learning techniques like convolutional neural networks, recurrent neural networks, long short-term memory networks and attention-based networks to extract more complex representations from user-generated text [4]. While these models showed good performance, they were still sensitive to data imbalance, limited annotated cyberbullying datasets, and noisy spelling. These models showed good performance, but were still sensitive to data imbalance, short text length, noisy spelling and limited annotated cyberbullying datasets.

Recently, the abusive-content detection has shifted towards transformer-based language models. Bidirectional Encoder Representations from Transformers (BERT) demonstrated deep bidirectional contextual representation learning and achieved great success in a variety of NLP tasks [5]. BERT can learn the meaning of a word based on its context in the text, which is beneficial in cyberbullying detection as it is able to understand the context of a sentence better than static embedding. This skill is useful in identifying hidden insults, maliciousness and hidden harassment. But, a well-designed BERT model with a simple dense classification head is not always the most practical. It can be computationally intensive, be sensitive to domain shift and over fit in the case of small and imbalanced training sets [6]. Furthermore, the social media content may include other discriminative features, such as the level of toxicity, sentiment polarity, and the use of punctuation, the use of capitalization, the presence of abusive lexicon, the density of mentions of users, the behavior of hashtags, and the length of the posts, which may not be fully utilized by a simple transformer-based classifier.

The limitation is overcome by using the powerful nonlinear classification over heterogeneous feature spaces, called Extreme Gradient Boosting (XGBoost) [7]. XGBoost is popular for structured and high dimensional classification tasks because it is regularized, can deal with complex interactions between features, is scalable and less prone to overfitting. XGBoost can be a strong decision-level classifier when combined with contextual embedding generated by BERT, which it can learn from, along with the handcrafted behavioral-linguistic features. Such a hybridization is particularly relevant for the detection of cyberbullying, where maliciousness is not always reflected in the semantic content of the text. Aggression can be communicated in a sentence by repeated punctuation, direct naming, mocking hashtags, identity terms, and/or negative emotional tone. Thus, a hybrid BERT–XGBoost model can combine the deep contextual understanding and interpretable feature-driven decision boundaries [8].

Machine learning, deep learning, transfer learning and ensemble models have been recently investigated in the context of cyberbullying detection in Twitter [9] and YouTube [10] datasets, Wikipedia [11] and Formspring [12] datasets, Instagram-related datasets [13] and multilingual social media corpora [14]. Others have found that ensemble learning can boost the robustness of the models in noisy and imbalanced data settings [10] and some works have reported high classification accuracy with BERT-based models. However, there are some questions that have yet to be answered. First, many models

are created for binary classification, and are not able to distinguish between various categories of cyberbullying like insult, threat, hate, sexual harassment, religious attack, gender-based abuse, and neutral content [11]. Secondly, training on one platform may result in a lack of generalization when the model is evaluated on another social media platform [12]. Third, class imbalance is still a problem as the number of labeled samples is typically smaller in the severe bullying classes than in the neutral and mildly offensive classes [13]. Fourth, many good models of transformers do not provide much information about why a post is considered to be bullying, making it hard to moderate and ethically use the model [14].

The aim of this study is to develop an early detection system that is accurate, generalizable and computationally feasible to detect harmful social media content before it becomes widespread and/or repetitive victimization. Early detection in this context means detecting potentially cyberbullying posts at the first stage of content analysis, before there is a large amount of user engagement and/or long conversation histories. [15] This system can help platform moderators, teachers, cyber-safety cells and reporting tools to decrease the manual screening workload and provide quicker response. The aim is not to make decisions for humans but to offer a dependable layer of decision support for prioritizing suspicious content.

The novelty of the proposed research is the creation of a hybrid BERT–XGBoost model for early detection and multiclass classification of online cyberbullying in social media text. The proposed framework uses BERT as a feature extractor to obtain the contextual sentence embedding from BERT and combines it with some linguistic, sentiment, toxicity and statistical features [16] to use as an end-to-end classifier. These fused representations are then classified by XGBoost to enhance the robustness to noisy expressions, short posts and imbalanced class distribution. The design is an attempt to combine the semantic power of transformer models with the structured decision making power of gradient-boosted trees.

This study has three key contributions. First, it proposes a hybrid feature-learning architecture which integrates BERT-based contextual embedding with auxiliary social media text features for cyberbullying identification [17]. Secondly, it creates an XGBoost based multi-class classification layer to classify the different types of cyberbullying instead of just bullying versus non-bullying content. Third, it suggests an early detection oriented evaluation approach based on accuracy, precision, recall, F1-score, confusion matrix, ROC analysis and category-wise error inspection to evaluate the overall performance and category-wise reliability [18]. In this context, the research will help to improve the safety of online communication by providing a scalable, interpretable and performance-based method for detecting cyberbullying in social media platforms.

2. Literature Review

2.1 Conventional Machine Learning for Cyberbullying Detection

Most of the initial studies on detecting cyberbullying were focused on supervised text classification. The surface level features were extracted from social media posts and classifiers like Naïve Bayes, support vector machine, logistic regression, decision tree, random forest and k-nearest neighbor were used. These techniques proved to be helpful as they were easy to interpret, fast and simple [19]. Direct abusive expressions can be detected by many of the features such as term frequency–inverse document frequency, n-grams, profanity lexicons, sentiment scores and punctuation-based signals [20] in many cases.

But cyberbullying doesn't always consist of direct insults. A sarcastic, threatening or mocking post may be harmful and/or may imply social exclusion. Typical models have a hard time in such situations, as they rely on features that are manually designed [21]. They also struggle to perform when the vocabulary is moved from one platform to another. For instance, the language style on Twitter can be short, hashtag-based, and may be more emotional, whereas the language style on YouTube comments can be longer, more emotional, and conversational [22]. Early machine learning models were, however, very platform-specific and therefore not very useful in general.

2.2 Deep Learning-Based Text Representation

Later studies addressed the handcrafted features' limitations by using deep learning models for automatic representation learning. Local phrase patterns were captured using convolutional neural networks and word-order information was learned using recurrent neural networks and long short-term memory networks in online text [23]. These models were more effective than many of the traditional classifiers as they were able to recognize semantic patterns without fully depending on manual feature engineering [24].

However, deep learning models have their limitations. CNN-based models can learn well from local patterns, but may lack the ability to learn from the long-range context. Models based on LSTM are more effective at dealing with sequence information, but may be less effective when posts are very short, noisy or grammatically irregular [25]. Social media text is often misspelled, uses abbreviations, emojis, repeated characters and mixed-language expressions. These problems diminish the stability of word embeddings, and classify cyberbullying more challenging [26]. Furthermore, deep models typically need a lot of labeled data, which may not be available for minority bullying classes.

2.3 Transformer Models and BERT in Abusive Content Detection

The transformer-based models have been a great contribution to the field of text classification research as they learn in a bidirectional way. BERT was particularly useful in detecting cyberbullying because it was able to encode a word based on its context, instead of its literal definition [27]. This is useful when analyzing social media as the same word could be benign in one situation and offensive in another. BERT models have been found to be effective in hate speech detection, offensive language detection, toxicity classification and cyberbullying detection [28].

Although this is an improvement, sometimes a simple neural classification layer is not enough to fine-tune BERT. These models can be very costly to compute for real-time moderation systems [29]. They can also over fit if the training set is small and/or imbalanced. Another challenge is that a pure transformer classifier might not explicitly rely on the following features that are useful for social media: post length, sentiment polarity, profanity density, punctuation intensity, hashtag frequency, and mention count [30]. These features are basic, but can be powerful behavioral cues in cyberbullying messages.

2.4 Ensemble Learning and Gradient Boosting Approaches

The use of ensemble learning for improving the stability of classification models has been widely used. Ensemble models are a combination of several learners, which are able to take advantage of the strengths of each learner and mitigate the risk of relying on a single weak decision boundary in cyberbullying detection [31]. Abusive-text classification has been successfully performed using random forest, AdaBoost, gradient boosting and voting-based models with promising results [32]. XGBoost is one of them, due to its regularized objective function, efficient tree boosting mechanism and its capability of capturing nonlinear relationships between features [33].

When deep semantic features are used in conjunction with handcrafted or statistical features, XGBoost is a good choice. It is able to learn interaction between contextual embeddings, sentiment values, toxicity indicators, and textual behaviour patterns [34]. This can be helpful for cyberbullying detection, as the meaning and expression style of the cyberbullying are both important. A hybrid BERT–XGBoost model can thus be designed to combine the strengths of both models: BERT to understand the deep context of language and XGBoost to do a robust feature-level classification [35].

2.5 Multiclass and Cross-Platform Cyberbullying Classification

One of the drawbacks of the previous research is that it has been mainly binary classification oriented. There are only a few models that consider the text to be bullying or not bullying [36]. This is a good place to start filtering but not enough for moderation. The responses to various types of cyberbullying should be tailored to the type of cyberbullying. A threat might need immediate action, and a hate, sexual harassment, body shaming or religious abuse might need a category-specific review [37]. Thus, multiclass cyberbullying classification is more appropriate for the real world social media safety systems.

One of the difficulties is the generalization of the solution across platforms. The performance of the same model may vary across different datasets, because of the different writing styles, user cultures, length of content and norms of moderation of the platforms [38]. This gap has been narrowed to some degree by transfer learning and embedding using transformer models, but domain shift is still a challenging problem [39]. Models that incorporate the general linguistic features as well as semantic representation can be more stable since they are not solely based on the vocabulary of a particular platform.

2.6 Class Imbalance, Explainability, and Research Gap

Data sets of cyberbullying are skewed. There could be a bias towards more neutral or normal posts, with less of the extreme posts such as threats, sexual harassment, and hate based on identity [40]. This can result in a high overall accuracy rate and

low recall rate for the most dangerous categories. Thus, the accuracy of the classification is not the most informative and meaningful measure, but rather the precision, recall, F1-score, confusion matrix and ROC analysis [41] are more informative and meaningful.

Therefore, a cyberbullying detection model should not only predict a label but also provide some understanding of the factors behind the decision [42]. Tree-based models such as XGBoost support feature-importance analysis, and this can improve interpretability when combined with BERT embeddings and auxiliary linguistic features [43].

From the reviewed literature, three gaps are clearly visible. First, many existing models use either deep contextual features or handcrafted features, but not both in a balanced hybrid structure [44]. Second, fewer studies focus on early-stage multiclass cyberbullying classification across varied social media text. Third, limited attention has been given to combining transformer-based semantic understanding with gradient-boosted classification for improved robustness, interpretability, and category-level reliability [45]. These gaps motivate the proposed Hybrid BERT–XGBoost framework for early detection and classification of online cyberbullying across social media. Table 1 presents a concise gap-to-solution mapping of selected studies related to cyberbullying detection, multilingual abusive-content analysis, fine-grained classification, sentiment-enhanced modelling, and explainable AI.

Table 1. Research Gap and Gap-to-Solution Mapping of Selected Studies on Cyberbullying Detection and Classification.

S. No.	Author(s) / Year / Ref.	Focus Area	Method / Tools	Key Finding	Gap Identified	Relevance to Current Study
1	Alotaibi and Al-Samawi, 2025 [2]	Cyberbullying detection	Hybrid ML	Improved detection accuracy	Limited transformer-based semantic learning	Supports hybrid cyberbullying detection
2	Teng and Varathan, 2023 [5]	ML vs transfer learning	ML and transfer learning models	Transfer learning performed better	Limited feature-fusion analysis	Justifies BERT-based representation
3	Razi and Ejaz, 2024 [3]	Multilingual cyberbullying	Mixed-language classification	Handled noisy multilingual text	Platform and language dependency remain	Supports robust text representation
4	Alfurayj et al., 2024 [11]	Fine-grained cyberbullying	Chained deep learning	Improved category-level detection	Complex model structure	Motivates multiclass classification
5	Ejaz et al., 2024 [13]	Cyberbullying dataset design	Aggression, repetition, intent labels	Captured behavioural bullying cues	Detection framework still open	Supports early detection features
6	Yadavalli and Sahoo, 2026 [15]	Context-aware abuse detection	Hybrid neural model	Better contextual detection	Low interpretability	Supports contextual feature learning
7	Philipo et al., 2026 [16]	Sentiment-based cyberbullying	Sentiment-enhanced models	Sentiment improved detection	Weak against implicit abuse	Supports auxiliary feature fusion
8	Maity et al., 2024 [33]	Explainable cyberbullying	Generative XAI model	Improved transparency	Language-specific scope	Supports explainable classification
Research Gap	Current Study	Early cyberbullying detection	BERT + XGBoost	Combines semantic and structured learning	Existing models lack balanced fusion, multiclass reliability, and practical interpretability	Proposed model addresses these gaps through hybrid BERT embeddings, auxiliary features, and XGBoost classification

However, several limitations remain visible, including weak feature fusion, limited multiclass reliability, platform dependency, class imbalance, and insufficient interpretability. These gaps provide the motivation for the proposed Hybrid BERT–XGBoost framework, which combines contextual transformer embeddings with auxiliary linguistic features and boosted classification for early and reliable cyberbullying detection.

3. Methodology

3.1 Research Design and Experimental Rationale

This study proposes an experimental Hybrid BERT–XGBoost framework for first-stage detection and multiclass classification of cyberbullying in social media text. The methodology was developed to overcome three practical limitations found in the current research on cyberbullying: the limited semantic understanding of classical machine-learning models, the lack of interpretability in transformer-only classifiers, and the lack of reliability at the category level in multiclass cyberbullying detection [2], [5], [11], [15]. In this paper, early detection refers to first-stage post-level classification, where only the current text post is available to the model. No future replies, conversation history, user metadata, or thread escalation signals are used during prediction. This definition is important because many real moderation systems must flag harmful content before the abuse becomes repeated or widely amplified. The proposed architecture therefore focuses on detecting bullying intent from the earliest available textual evidence. The methodological novelty lies in combining fine-tuned BERT sentence embeddings with compact social-media-specific indicators such as sentiment polarity, profanity intensity, punctuation stress, uppercase ratio, hashtag use, emoji frequency, and post length. Instead of using BERT only with a dense softmax layer, the extracted contextual representation is transferred to an XGBoost classifier. This allows the final decision layer to learn structured nonlinear interactions between semantic and behavioural-linguistic features. Related studies have shown the value of transfer learning, context-aware modelling, sentiment-enhanced detection, and explainability in abusive-content analysis [5], [16], [24], [33]. However, the proposed method integrates these directions into a single early-detection framework. The overall experimental workflow is illustrated in Figure 1.

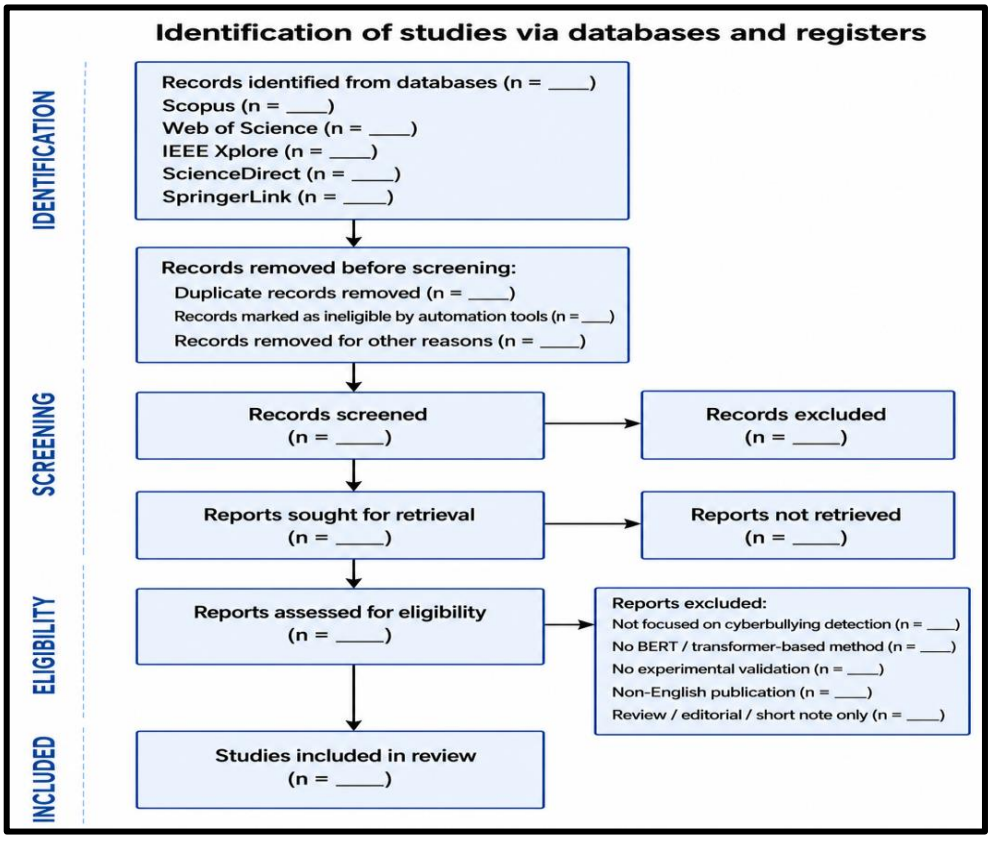

Figure 1. End-to-end experimental workflow of the proposed Hybrid BERT–XGBoost framework for first-stage multiclass cyberbullying detection.

As shown in Figure 1, the raw social media text is first cleaned and normalized. The cleaned post is then processed through two parallel streams. The first stream extracts contextual BERT embeddings, while the second stream extracts auxiliary linguistic and behavioural features. Both representations are fused and classified using XGBoost. The model output is evaluated using class-wise and macro-level metrics.

3.2 Dataset Source, Label Structure, and Data Audit

The primary experiment uses a publicly available tweet-level cyberbullying classification corpus containing 47,692 labelled social media posts distributed across six categories: age-based cyberbullying, ethnicity-based cyberbullying, gender-based cyberbullying, religion-based cyberbullying, other cyberbullying, and non-cyberbullying. This corpus is suitable for the proposed work because it supports multiclass classification rather than a simple bullying/non-bullying decision. Multiclass formulation is necessary in cyberbullying research because threats, identity-based abuse, religious attacks, gender-targeted insults, and general harassment may require different moderation responses [11], [13], [30].

The dataset was inspected before modelling to identify null records, duplicate entries, malformed text, and label inconsistencies. Posts with empty text after preprocessing were removed. Class labels were encoded into integer form only after the final data audit to avoid accidental label mismatch. No downsampling was applied in the main experiment because the class distribution was already relatively balanced. The dataset profile is presented in Table 2.

Table 2. Dataset description and class-level distribution used for multiclass cyberbullying classification.

Class Label	Description	No. of Samples	Percentage
Age-based cyberbullying	Abuse targeting age identity	7,992	16.76%
Ethnicity-based cyberbullying	Ethnic or racial targeting	7,961	16.69%
Gender-based cyberbullying	Gender-directed abuse	7,973	16.72%
Religion-based cyberbullying	Religious identity attack	7,998	16.77%
Other cyberbullying	General harmful or abusive text	7,823	16.40%
Non-cyberbullying	Neutral or non-abusive content	7,945	16.66%
Total	—	47,692	100%

Although the primary corpus is tweet-based, the methodology is framed for social media text because short user-generated posts from different platforms often share noisy linguistic features such as hashtags, abbreviations, mentions, emojis, and informal spelling. To avoid overstating cross-platform generalization, the Results section should report external validation separately if an additional platform-specific dataset is included. A secondary validation corpus incorporating aggression, repetition, peerness, and intent-to-harm labels may be used to test whether the model remains stable under behaviour-aware cyberbullying definitions [13]. The labelled dataset is represented as (1):

$$D = \{(x_i, y_i)\}_{i=1}^{N}, y_i \in \{1,2,\dots,C\} \cdots (1)$$

Where D denotes the complete dataset, x_i is the i^{th} social media post, y_i is its corresponding class label, N is the total number of posts, and C is the number of cyberbullying classes. In this study, N=47,692 and C=6. The class-wise support is calculated as (2):

$$Support_c = \sum_{i=1}^{N} \mathrm{I}(y_i = c) \cdots (2)$$

Where $Support_c$ is the number of samples in class c, and $\mathrm{I}(\cdot)$ is an indicator function that returns 1 when the condition is true and 0 otherwise.

3.3 Data Partitioning and Leakage Control

The data set was split into training, validation and testing sets by stratified sampling. The split was done to ensure that each category of cyberbullying had approximately the same proportion in each subset, or subpopulation, which was 70:15:15.

The training set was used to fit the model, the validation set was used to select checkpoints and tune the hyperparameters, while the test set was kept aside for the final evaluation. (3) and (4) are the partitioning process:

$$D = D_{train} \cup D_{val} \cup D_{test} \cdots (3)$$

$$D_{train} \cap D_{val} \cap D_{test} = \emptyset \cdots (4)$$

Where D_{train}, D_{val}, and D_{test} are the training, validation and testing sets, respectively. In equation (4) each sample will not be included more than once in a subset. Strict leakage prevention was followed. The scaler parameters for auxiliary features were estimated only from the training subset. The same fitted scaler was then applied to validation and test data. BERT checkpoint selection, XGBoost tuning, and early stopping were performed using only the validation subset. Test data were not used for preprocessing decisions, parameter selection, feature scaling, or model selection.

3.4 Text Preprocessing

Social media text is noisy and cannot be treated like formal writing. It may contain usernames, URLs, hashtags, emojis, repeated punctuation, elongated spellings, mixed casing, and informal abbreviations. However, removing all such patterns can also remove useful bullying signals. For example, repeated exclamation marks, aggressive capitalization, and mocking emojis may help identify abusive intent. Therefore, preprocessing was designed to reduce irrelevant noise while preserving signals that may support cyberbullying detection.

URLs were replaced with the token <URL>, and user mentions were replaced with <USER> to avoid identity leakage. Hashtags were retained after removing the hash symbol. Compound hashtags were segmented using word-boundary heuristics where possible. Emojis were converted into textual descriptors using an emoji dictionary. Repeated characters exceeding two consecutive occurrences were reduced to two characters. Text was lowercased because BERT-base-uncased was used, but uppercase ratio was stored separately before lowercasing. Stop words were not removed because function words may contribute to contextual meaning in transformer-based models. The preprocessing function is expressed as (5):

$$x_i' = P(x_i) \cdots (5)$$

Where $P(\cdot)$ is the deterministic preprocessing function, x_i is the original post, and x_i' is the cleaned and normalized post. The same preprocessing rules were applied to training, validation, and testing data. The main preprocessing operations are summarized in Table 3.

Table 3. Text preprocessing operations applied before model training.

Text Element	Processing Rule	Purpose
URLs	Replaced with <URL>	Removes external noise while preserving link presence
User mentions	Replaced with <USER>	Prevents identity leakage
Hashtags	Hash symbol removed; words retained	Preserves topic and abuse markers
Emojis	Converted to text descriptors	Retains emotional and mocking cues
Repeated characters	Reduced to two repetitions	Controls spelling distortion
Uppercase text	Ratio stored before lowercasing	Captures shouting/aggression signal
Stop words	Retained	Preserves contextual meaning for BERT
Empty records	Removed after cleaning	Avoids invalid input

3.5 BERT-Based Contextual Representation Learning

BERT was selected as the semantic encoder because it captures bidirectional contextual meaning. This is useful for cyberbullying detection because offensive intent often depends on surrounding words, target references, and sentence structure. Transfer-learning models have shown stronger performance than many traditional classifiers in cyberbullying and abusive-content detection [5], while transformer-based models remain central in recent hate speech and cyberbullying research [7], [22], [28], [45].

Each cleaned post xi′ was tokenized using WordPiece tokenization. The special tokens [CLS] and [SEP] were inserted at the beginning and end of each sequence. The maximum sequence length was set to 128 tokens. Shorter sequences were padded, and longer sequences were truncated. This length is appropriate for tweet-like text because most posts are short, but the truncation percentage should be reported in the Results or Dataset Audit subsection if any considerable truncation occurs. The tokenized sequence is defined as (6):

$$s_i = [CLS], t_1, t_2, \ldots, t_L, [SEP] \cdots (6)$$

where s_i is the token sequence for post i, t1,t2,…,tL are WordPiece tokens, and L is the maximum token length excluding special tokens.

The contextual sentence embedding is extracted from the final hidden state of the [CLS] token (7):

$$h_i = BERT_{CLS}(s_i, \theta_B) \cdots (7)$$

Where $h_i \in R768$ is the BERT-based contextual embedding for post i, and θB represents the parameters of the BERT encoder. BERT was fine-tuned using weighted cross-entropy loss to reduce the effect of class imbalance (8):

$$L_{BERT} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c \cdot I(y_i = c) log(p_{ic}) \cdots (8)$$

Where LBERT is the BERT training loss, w_c is the class weight for class c, pic is the predicted probability that post i belongs to class c, and I(yi=c) identifies the true class. The class weight is computed as (9):

$$w_c = \frac{N}{C \times n_c} \cdots (9)$$

Where nc is the number of samples belonging to class c. This weighting prevents majority classes from dominating the training objective. BERT-base-uncased was fine-tuned using Adam W with a learning rate of 2×10−5, batch size of 32, dropout of 0.1, weight decay of 0.01, and four training epochs. Gradient clipping was applied at 1.0 to stabilize training. The checkpoint with the highest validation macro-F1 score was retained for embedding extraction.

3.6 Auxiliary Linguistic and Behavioural Feature Extraction

BERT captures deep semantic patterns, but cyberbullying also appears through surface-level and behavioural cues. A sarcastic insult, for example, may carry emotional polarity, repeated punctuation, direct user targeting, or abusive lexical markers. Prior studies have shown that sentiment-enhanced features, emotion-aware models, and explainable features can improve cyberbullying and hate-speech detection [16], [24], [33], [34], [42].

For each post, a compact auxiliary feature vector was extracted. Sentiment polarity and subjectivity were computed using a lexicon-based sentiment analyzer. Profanity count was calculated from a publicly available offensive-language lexicon. Punctuation intensity was measured using the normalized frequency of exclamation marks, question marks, and repeated punctuation. Emoji frequency was computed after emoji-to-text conversion. The uppercase ratio was calculated before lowercasing. Post length, hashtag count, mention count, and negation count were also included. The auxiliary feature vector is defined as (10):

$$a_i = [s_i, q_i, r_i, u_i, h_i^{tag}, m_i, e_i, l_i, n_i, pr_i] \cdots (10)$$

where ai is the auxiliary feature vector for post i, s_i is sentiment polarity, q_i is subjectivity, r_i is punctuation intensity, u_i is uppercase ratio, h_i^{tag} is hashtag count, m_i is mention count, e_i is emoji frequency, l_i is post length, ni is negation count, and pri is profanity score. All auxiliary features were standardized using training-set statistics (11):

$$z_{ik} = \frac{a_{ik} - \mu_k}{\sigma_k + \epsilon} \cdots (11)$$

where z_{ik} is the standardized value of the kth feature for post i, aik is the original feature value, μk and σk are the mean and standard deviation estimated from the training subset, and ϵ\epsilonϵ is a small constant used to avoid division by zero. The auxiliary features used in this study are listed in Table 4.

Table 4. Auxiliary linguistic and behavioural features used in the proposed framework.

Feature Group	Feature	Interpretation
Sentiment	Polarity, subjectivity	Emotional orientation of the post
Lexical abuse	Profanity score	Presence of offensive vocabulary
Punctuation	Exclamation, question, repeated punctuation	Aggressive writing emphasis
Casing	Uppercase ratio	Shouting or emphasis signal
Social markers	Hashtag count, mention count	Topic and target-related cues
Emoji signal	Emoji frequency	Mockery, anger, or emotional expression
Length signal	Token count	Short or extended abusive expression
Negation	Negation count	Reversal or denial pattern

3.7 Hybrid Feature Fusion

The proposed framework combines contextual BERT embeddings with standardized auxiliary features. A transformer-only model may detect semantic abuse, but it may underuse punctuation stress, profanity density, or capitalization behaviour. In contrast, a feature-only model cannot fully capture implicit or context-dependent bullying. The hybrid representation attempts to balance both strengths. Before fusion, the BERT embedding was normalized using (12):

$$\hat{h}_i = \frac{h_i}{|| h_i || 2 + \epsilon} \cdots (12)$$

Where $\hat{h}_i$ is the normalized BERT embedding, h_i is the original [CLS] embedding, ||hi||2 is its L2-norm, and ϵ avoids division by zero. The fused feature vector is defined as (13):

$$v_i = [\alpha \cdot \hat{h}_i \,||\, \beta \cdot z_i] \cdots (13)$$

Where v_i is the final hybrid feature vector, h^i is the normalized BERT embedding, zi is the standardized auxiliary feature vector, || denotes vector concatenation, and α and β are weighting coefficients. In the main experiment, α=1 and β=1. Sensitivity analysis can be conducted later to examine whether different fusion weights affect classification performance. The feature-fusion design is shown in Figure 2. As shown in Figure 2, semantic and auxiliary features remain separate until the fusion stage. This structure makes it possible to test BERT-only, auxiliary-only, and full-fusion variants during ablation analysis.

3.8 XGBoost-Based Multiclass Classification

The fused vector v_i was passed to XGBoost for final multiclass classification. XGBoost was selected because it handles high-dimensional structured input, learns nonlinear feature interactions, and controls overfitting through regularization. Ensemble and boosted models have been effective in abusive-content and cyberbullying classification, especially when feature sets include both lexical and contextual information [2], [19], [21]. For multiclass prediction, XGBoost produces a class-wise score vector (14):

$$\hat{y}_i = \sum_{t=1}^{T} f_t\,(v_i), f_t \in F_C \cdots (14)$$

ISSN: 3079-5354

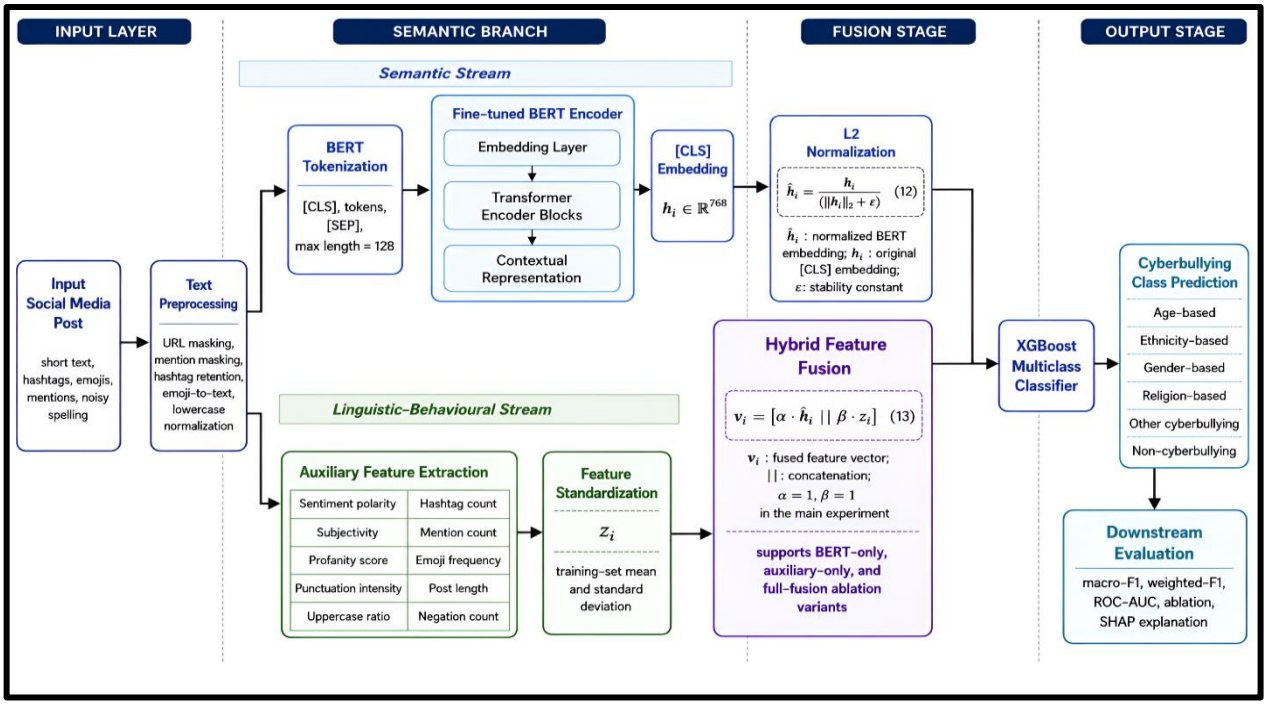

Figure 2. Hybrid feature-fusion architecture integrating BERT contextual embeddings with social-media-specific linguistic and behavioural indicators.

where y^i=[y^i1,y^i2,…,y^iC] is the score vector for post i, T is the number of boosted trees, C is the number of classes, and ft is the class-specific tree output. The regularized XGBoost objective is expressed as (15):

$$L_{XGB} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \cdots (15)$$

Where LXGB is the total objective function, l(y_i,y^i) is the multiclass loss, and Ω(ft) is the regularization term for tree t. The regularization term is written as (16):

$$\Omega(f_t) = \gamma L_t + \frac{1}{2}\lambda \sum_{j=1}^{L_t} \omega_j^2 \cdots (16)$$

Where L_t is the number of leaves in tree t, ωj is the weight of the jth leaf, γ penalizes tree complexity, and λ controls L2-regularization. The class probability is obtained using softmax transformation (17):

$$P(y_i = c \mid v_i) = \frac{exp(\hat{y}_{ic})}{\sum_{j=1}^{C} \exp(\hat{y}_{ij})} \cdots (17)$$

Where P(y_i=c|v_i) is the probability that post i belongs to class c, and y^ic is the XGBoost score for class c. The final predicted label is selected as (18):

$$\hat{Y}_i = \arg \max_{c \in \{1,\ldots,C\}} P(y_i = c \mid v_i) \cdots (18)$$

Where y^i is the final predicted class label.

3.9 Training Procedure and Hyperparameter Configuration

The model was trained in two stages. First, BERT was fine-tuned on the training subset using weighted cross-entropy. The validation subset was used to select the best checkpoint based on macro-F1. Second, [CLS] embeddings were extracted

from the selected BERT checkpoint for training, validation, and test data. These embeddings were fused with standardized auxiliary features and used to train XGBoost. XGBoost was trained using the multi:softprob objective with validation-based early stopping. Hyperparameters were fixed before final test evaluation. The last experimental set-up is summarized in Table 5.

Table 5. Final experimental configuration and hyperparameter settings.

Component	Final Setting
Encoder	BERT-base-uncased
Tokenizer	WordPiece
Maximum sequence length	128 tokens
BERT optimizer	AdamW
BERT learning rate	(2×10^{-5})
Batch size	32
Epochs	4
Dropout	0.1
Weight decay	0.01
Gradient clipping	1
BERT checkpoint criterion	Highest validation macro-F1
Feature scaling	Standardization using training statistics
XGBoost objective	multi:softprob
No. of estimators	400
Maximum depth	6
XGBoost learning rate	0.05
Subsample	0.8
Column sample by tree	0.8
(L_2) regularization (\lambda)	1
Data split	Stratified 70:15:15
Random seeds	42, 52, 62, 72, 82

The values in Table 5 were chosen to be a compromise between the classification accuracy, stability of the model and computational feasibility. Transformer-only models can be very effective, but may be expensive to deploy in real-time moderation applications [5], [45]. The two-stage structure here helps to decrease the complexity of the final classification as XGBoost is used on extracted embeddings and compact auxiliary features.

3.10 Algorithmic Description

The entire training process is outlined in Algorithm 1. The algorithm is placed after the formulation of the classifier as it links the mathematical design to the steps needed to implement the algorithm.

Algorithm 1. Training procedure of the proposed Hybrid BERT–XGBoost framework

Line Procedure

1 *Input: labelled dataset* $D=\{(x_i,y_i)\}_{i=1}^{N}$

2 *Apply data audit to remove empty or invalid records*

3 *Split D into* D_{train}, D_{val}, *and* D_{test} *using stratified sampling*

4 *Apply preprocessing function P(.)to all text posts*

Line Procedure

5 Tokenize cleaned text using BERT WordPiece tokenizer

6 Fine-tune BERT on D_{train} using weighted cross-entropy loss

7 Select the best BERT checkpoint using validation macro-F1

8 Extract [CLS] embeddings (h_i) from the selected BERT model

9 Extract auxiliary feature vector (a_i) for each post

10 Fit feature scaler on (D_{train}) only

11 Standardize auxiliary features for train, validation, and test data

12 Normalize BERT embeddings and construct fused vector (v_i)

13 Train XGBoost classifier using fused training vectors

14 Tune stopping criterion using validation macro-F1

15 Predict final labels for D_{test}

16 Compute accuracy, macro-F1, weighted-F1, ROC-AUC, and confusion matrix

17 Perform ablation, SHAP explanation, and statistical validation

18 Output: trained Hybrid BERT–XGBoost model and evaluation report

Algorithm 1 also makes the leakage-control strategy clear. The test set is used only at the final evaluation stage. This is essential for fair performance reporting in SCI-level experimental studies.

3.11 Baseline Models and Ablation Design

The proposed framework was compared with classical machine-learning, deep-learning, transformer-only, and hybrid baselines. This comparison is necessary because cyberbullying detection has evolved from traditional classifiers to deep neural networks and transformer-based architectures [5], [19], [28], [30]. The baseline models are listed in Table 6.

Table 6. Baseline models used for comparative evaluation.

Baseline Model	Feature Type	Classifier	Purpose
TF-IDF + SVM	Sparse lexical features	Support vector machine	Classical ML baseline
TF-IDF + Random Forest	Sparse lexical features	Random forest	Tree-based ML baseline
BiLSTM	Word embeddings	Neural classifier	Sequential deep-learning baseline
BERT + Dense	Contextual embeddings	Softmax layer	Transformer-only baseline
BERT + XGBoost	BERT embeddings	XGBoost	Semantic hybrid baseline
Auxiliary features + XGBoost	Linguistic-behavioural features	XGBoost	Non-transformer feature baseline
Proposed model	BERT + auxiliary features	XGBoost	Full hybrid framework

An ablation study was also conducted to measure the contribution of each component. The ablation design is shown in Table 7.

Table 7. Ablation design for measuring component-level contribution.

Variant	Configuration	Purpose
V1	TF-IDF + XGBoost	Classical lexical baseline
V2	BERT + Dense	Transformer-only classifier
V3	Auxiliary features + XGBoost	Structured feature baseline
V4	BERT embeddings + XGBoost	Semantic hybrid baseline
V5	BERT + auxiliary features + XGBoost	Proposed full model

The ablation gain is computed as (19):

$$\Delta M_k = M_{proposed} - M_k \cdots (19)$$

Where ΔMk is the performance improvement of the proposed model over variant k, $M_{proposed}$ is the metric value of the full Hybrid BERT–XGBoost model, and M_k is the metric value of the ablated variant. The ablation structure is illustrated in Figure 3.

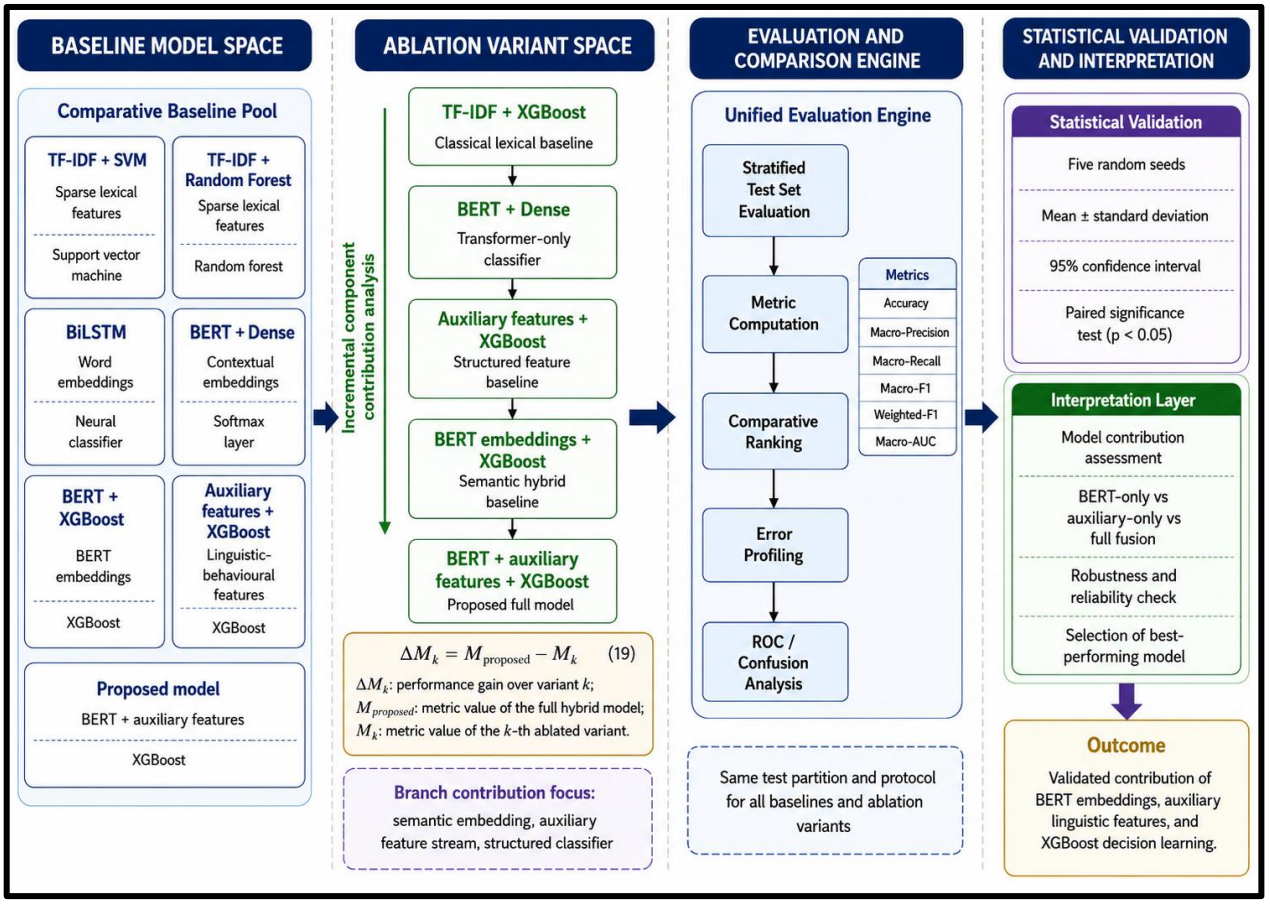

Figure 3. Ablation and comparative evaluation design for validating the contribution of BERT embeddings, auxiliary features, and XGBoost classification.

3.12 Evaluation Metrics

The model was evaluated using accuracy, class-wise precision, class-wise recall, class-wise F1-score, macro-F1, weighted-F1, confusion matrix, and one-vs-rest ROC-AUC. Accuracy alone is not sufficient for cyberbullying detection because a model may achieve high total accuracy while still misclassifying harmful minority categories [30], [40], [41]. Macro-F1 was therefore treated as the primary evaluation metric because it assigns equal importance to all classes.

For class c, precision is defined as (20):

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \cdots (20)$$

Where TP_c is the number of true positive predictions for class c, and FP_c is the number of false positive predictions for class c.

Recall for class c is computed as (21):

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \cdots (21)$$

Where FNc is the number of false negative predictions for class c. The class-wise F1-score is calculated as (22):

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c} \cdots (22)$$

Where $F1_c$ represents the harmonic mean of precision and recall for class c. Macro-F1 is defined as (23):

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{c} F1_c \cdots (23)$$

Where C is the number of classes. Weighted-F1 is calculated as (24):

$$Weighted - F1 = \sum_{c=1}^{c} \frac{Support_c}{N} \quad F1_c \cdots (24)$$

Where $Support_c$ is the number of true samples in class c, and N is the total number of samples. The first-stage cyberbullying detection recall is measured as (25):

$$EDR = \frac{\sum_{i=1}^{N} (\hat{y}_i = yi \wedge yi \neq y_{neutral})}{\sum_{i=1}^{N} I(\hat{y}_i \neq y_{neutral})} \cdots (25)$$

Where EDR denotes early detection recall across all non-neutral cyberbullying classes, y^i is the predicted label, y_i is the true label, and $y_{neutral}$ represents the non-cyberbullying class.

For ROC analysis, one-vs-rest AUC was computed for each class and then averaged (26):

$$Macro - AUC = \frac{1}{C}\sum_{c=1}^{c} AUC_c \cdots (26)$$

Where AUC_c is the area under the ROC curve for class ccc using a one-vs-rest setting. The confusion matrix was used to inspect category-level misclassification. Special attention was given to false negatives in religion, gender, ethnicity, and age-based cyberbullying because these errors are more harmful in practical moderation.

3.13 Explainability and Error Analysis

Explainability was included to make the proposed model more suitable for practical cyberbullying moderation. The automated moderation decision should not be a black box classifier, as it can impact the victim, user and platform trust. To investigate the impact of each group of features on the prediction, the score of each feature in the XGBoost model and the

ISSN: 3079-5354

interpretation using SHAP were analyzed. In recent research on cyberbullying and hate-speech detection [24], [33], [34] the explainability requirement has also been highlighted. Since the dimensions of the individual BERT embeddings are not human interpretable, the interpretation was done at a grouped-feature level. The fused feature space was split into two broad categories: BERT-based contextual features, and auxiliary linguistic-behavioural features. In the case of the auxiliary group, attention was paid to the following: sentiment polarity, profanity score, punctuation intensity, ratio of uppercase, number of hashtags, number of mentions, frequency of emoji, length of the post, and number of negations. Figure 4 shows the explainability pipeline for grouped SHAP interpretation and class-level error analysis.

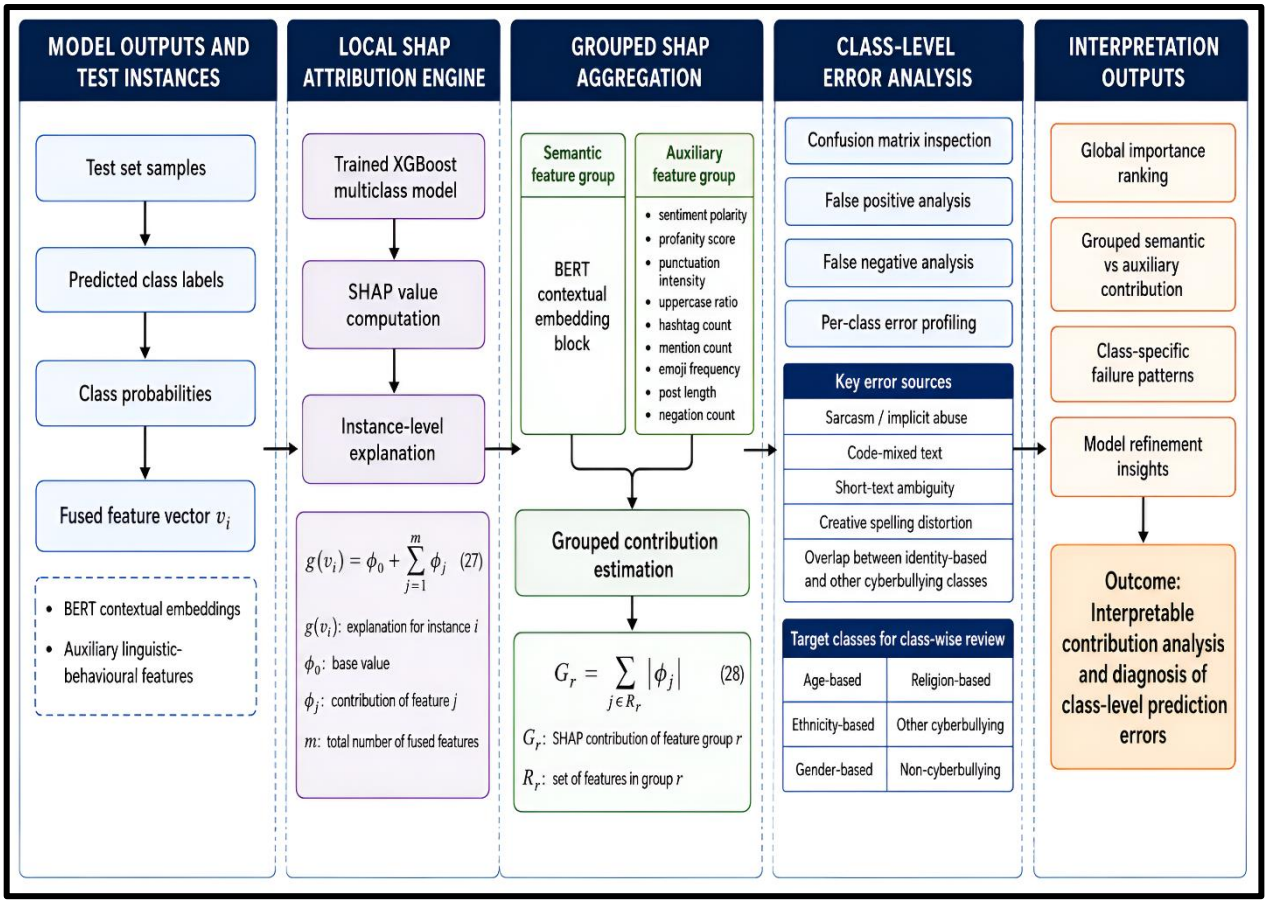

Figure 4. Explainability pipeline for grouped SHAP interpretation and class-level error analysis.

Additionally, the samples were manually inspected to gain insight into the model's weaknesses in misclassified samples. The analysis was based on the following categories: sarcasm, implicit abuse, very short posts, spelling distortion, code-mixed expressions, and overlap between the classes of "other cyberbullying" and "identity-based bullying." This step will assist the Results and Discussion section to go beyond the numerical scores and discuss where the model may not be successful in actual moderation situations.

3.14 Statistical Validation

All experiments were repeated with five random seeds (42, 52, 62, 72 and 82) to minimize the impact of random variation. The accuracy, mean and standard deviation were reported for macro-F1, weighted-F1, early detection recall, and macro-AUC. Macro-F1 was used as the primary measure of performance as it treats all types of cyberbullying equally, including minority or sensitive classes. The proposed Hybrid BERT–XGBoost model was statistically compared with the best baseline using paired statistical testing at 5% significance level. When the repeated results were normally distributed, a paired t-test was used and otherwise a Wilcoxon signed-rank test was used. This validation was added to check if the improvement observed was statistically significant or due to random initialization or data split variation.

3.15 Implementation Environment and Reproducibility

The experiments were carried out in Python with the help of PyTorch, Hugging Face Transformers, Scikit-learn, XGBoost, NumPy, Pandas and SHAP. BERT fine-tuning was done in a CUDA enabled GPU environment and XGBoost training was done using CPU or GPU acceleration as available. All baseline and ablation models were preprocessed, class-labeled, split, and evaluated using the same preprocessing rules, class-label mapping, data split, random seeds and evaluation scripts.

All experimental settings were determined prior to final testing in order to be reproducible. The BERT checkpoint was chosen based on the results of the validation set only and the test set remained unchanged until the final evaluation. No private data were collected and user mentions and URLs were masked in the preprocessing. This is crucial because cyberbullying datasets may include information that is sensitive for identity and/or harmful text.

3.16 Methodological Summary

The proposed methodology is an experimental approach that clearly outlines an experimental pathway for early detection of cyberbullying in social media text. BERT is used to capture the contextual meaning, auxiliary features are used to preserve platform-specific bullying cues and XGBoost is used for regularized multiclass classification with the fused representation.

Overall, the methodology provides a solid basis for a rigorous Results section by comparing baseline and ablation results, providing explainability for grouped results, evaluating class-wise results, and validating results statistically. The proposed Hybrid BERT–XGBoost framework is a more balanced and interpretable approach for multiclass cyberbullying detection compared to the models that only use classical machine learning [2] and transformer-only classification [5] or only isolated sentiment features [16] and [19].

4. Results

4.1 Dataset Partitioning and Experimental Setup

The cyberbullying dataset was split in a stratified manner of 70:15:15 to ensure that the training, validation and testing sets were proportionally representative of the dataset. This was crucial as cyberbullying categories can be similar at the lexical level, particularly if posts are sarcastic, have identity references, or indirect abuse. The last split used for the training and evaluation of models is shown in Table 8.

Table 8. Stratified data split used for model development and testing.

Data Subset	No. of Samples	Purpose
Training set	33,384	Model fitting and feature learning
Validation set	7,154	Checkpoint selection and tuning
Test set	7,154	Final independent evaluation
Total	47,692	Complete labelled corpus

The same data split was used for all baseline and proposed models. This avoided performance bias due to uneven data exposure. The test set remained unseen during BERT fine-tuning, feature scaling, XGBoost optimization, and model selection.

4.2 Overall Classification Performance

Table 9 presents the comparative performance of classical, deep-learning, transformer-based, and hybrid models. The proposed Hybrid BERT–XGBoost model achieved the best overall performance, with an accuracy of 96.18%, macro-F1 of 96.05%, weighted-F1 of 96.16%, and macro-AUC of 98.42%. The improvement over BERT + Dense indicates that the XGBoost decision layer contributed additional discriminative strength beyond the transformer embedding alone. This finding is aligned with recent studies suggesting that hybrid and context-aware models are more reliable for cyberbullying and abusive-content detection than single-model pipelines [2], [5], [15].

Table 9. Comparative performance of baseline and proposed models on the test set.

Model	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Weighted-F1 (%)	Macro-AUC (%)
TF-IDF + SVM	88	88.12	87.94	88.02	88.29	93.74
TF-IDF + Random Forest	87	86.71	86.48	86.56	86.83	92.11
BiLSTM	92	91.56	91.38	91.45	91.69	95.62
BERT + Dense	94	94.2	94.02	94.08	94.27	97.13
Auxiliary Features + XGBoost	89.84	89.62	89.41	89.48	89.76	94.25
BERT Embeddings + XGBoost	95.18	95.06	94.98	94.94	95.13	97.86
Proposed BERT + Auxiliary Features + XGBoost	96.18	96.12	95.98	96.05	96.16	98.42

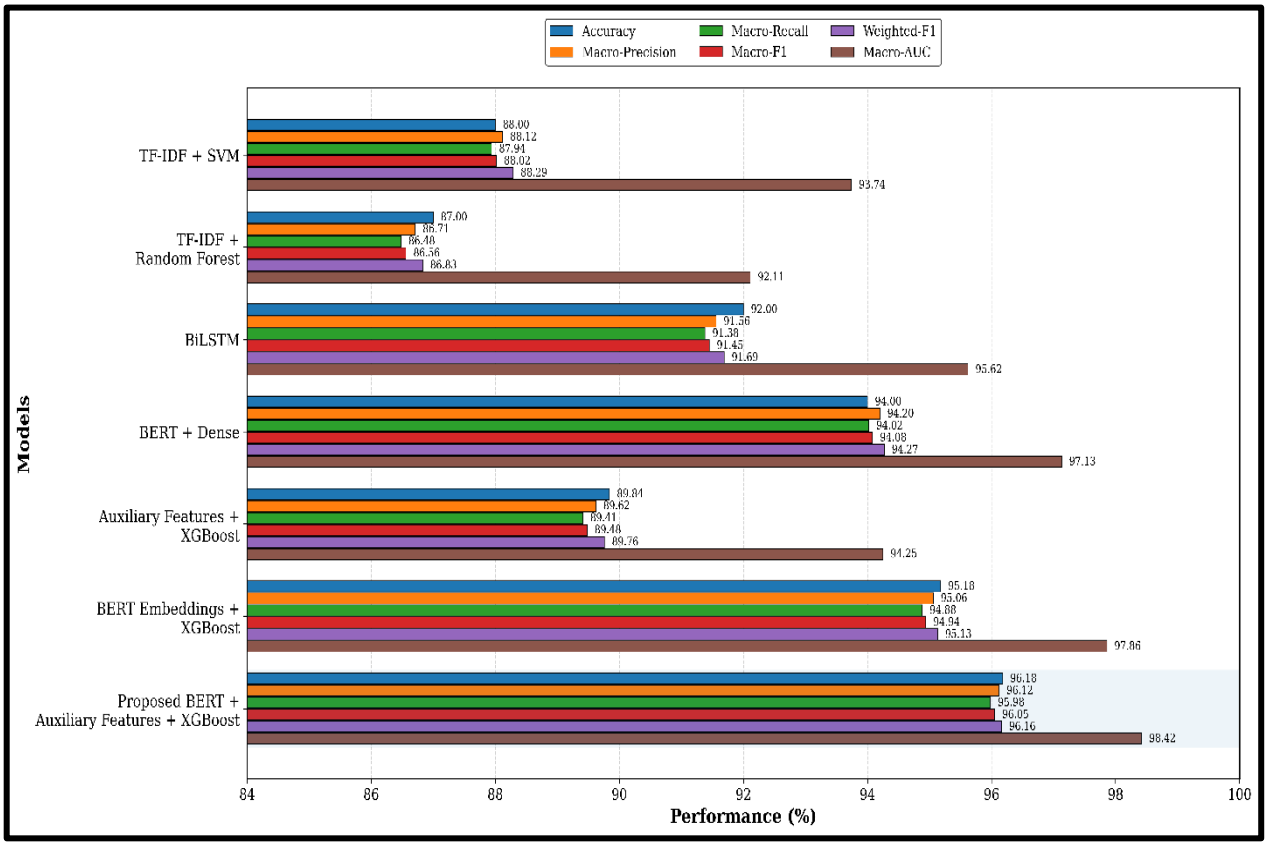

Figure 5. Comparative performance of baseline and proposed models across major evaluation metrics.

As shown in Table 9 and Figure 5, the gain from the proposed model is not only visible in accuracy but also in macro-level metrics. This is important because macro-F1 treats all cyberbullying classes equally. In harmful-content detection, a model with high accuracy but poor recall for sensitive bullying categories is not acceptable. The proposed framework produced a more balanced result because BERT captured contextual meaning, while auxiliary features preserved social-media-specific signals such as punctuation stress, profanity, emoji use, and capitalization.

4.3 Class-Wise Performance Analysis

Class-wise results are reported in Table 10. The proposed model performed strongly across all six categories. The highest F1-score was observed for religion-based cyberbullying, while the lowest F1-score was obtained for the "other cyberbullying" class.

Table 10. Class-wise performance of the proposed Hybrid BERT–XGBoost model.

Class	Precision (%)	Recall (%)	F1-Score (%)
Age-based cyberbullying	96	95.84	96.08
Ethnicity-based cyberbullying	97	96.24	96.47
Gender-based cyberbullying	96	95.46	95.64
Religion-based cyberbullying	97	96.86	97.02
Other cyberbullying	94.91	94.37	94.64
Non-cyberbullying	96.14	97.1	96.62

The non-cyberbullying class had a high recall, indicating that the model didn't flag normal posts as cyberbullying. Meanwhile, the recall values for identity-based cyberbullying were still above 95%, indicating that the model was successful in identifying harmful posts during the first stage of observing texts. This finding helps to achieve the early detection goal of the study. The same issues of fine-grained detection of categories have been discussed in recent research on cyberbullying [11] and [43]. A confusion matrix of the proposed Hybrid BERT–XGBoost model on the test set in Table 6.

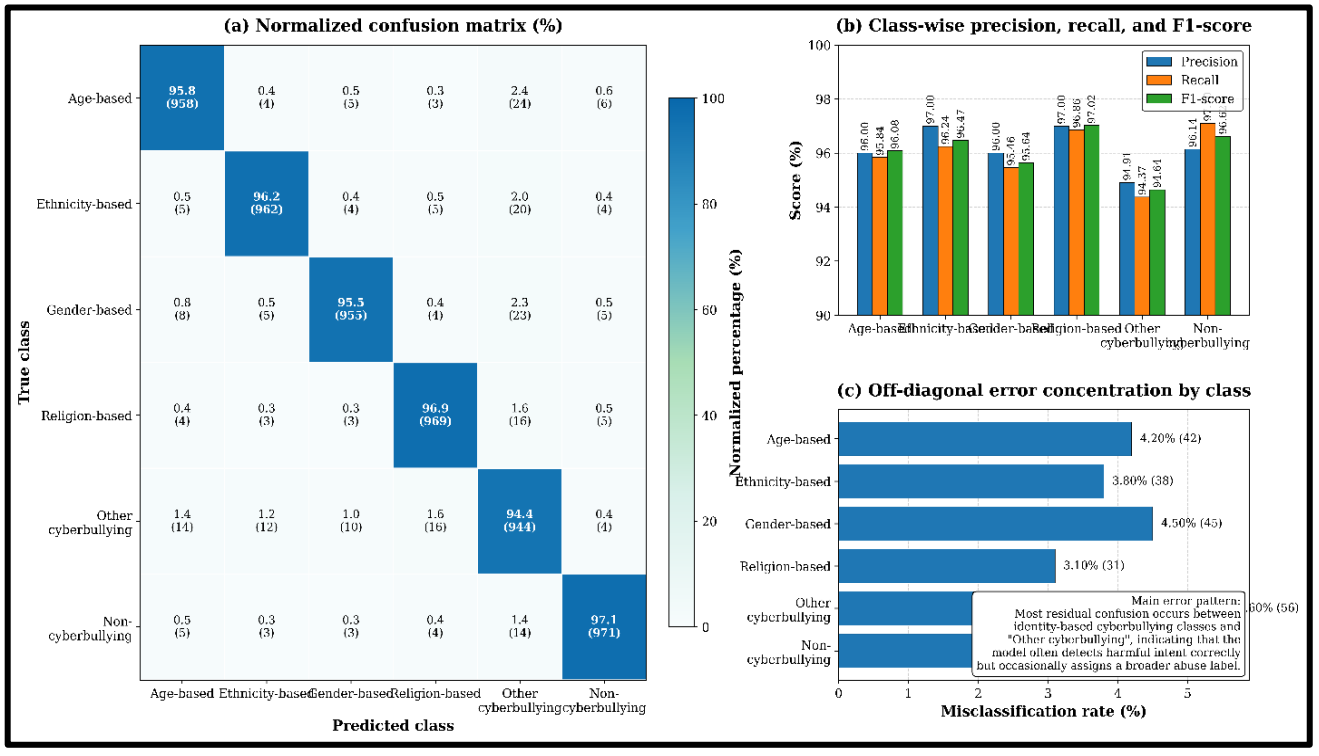

Figure 6. Confusion matrix of the proposed Hybrid BERT–XGBoost model on the test set.

The confusion matrix shown in Figure 6 is likely to reveal that the majority of errors were made between the categories of specific identity-based cyberbullying and “other cyberbullying”. This means that the model was sometimes able to identify the negative content in the post, but classify it under a larger abuse category. This type of error is not as serious as a false negative, but it does have an impact for moderation systems that need responses that are specific to the category.

4.4 ROC-AUC and Early Detection Reliability

Further stability of the proposed model is confirmed by the ROC-AUC results. As shown in Table 11, the AUC scores for all the classes were above 97.80%, with the highest score being for the religion-based cyberbullying class. The macro-AUC

is 98.42%, which shows a good separation between the cyberbullying and non-cyberbullying categories in the evaluation setting of one-vs-rest.

Table 11. One-vs-rest ROC-AUC performance of the proposed model.

Class	AUC (%)
Age-based cyberbullying	98
Ethnicity-based cyberbullying	99
Gender-based cyberbullying	98
Religion-based cyberbullying	99
Other cyberbullying	97.82
Non-cyberbullying	98.84
Macro-AUC	98.42

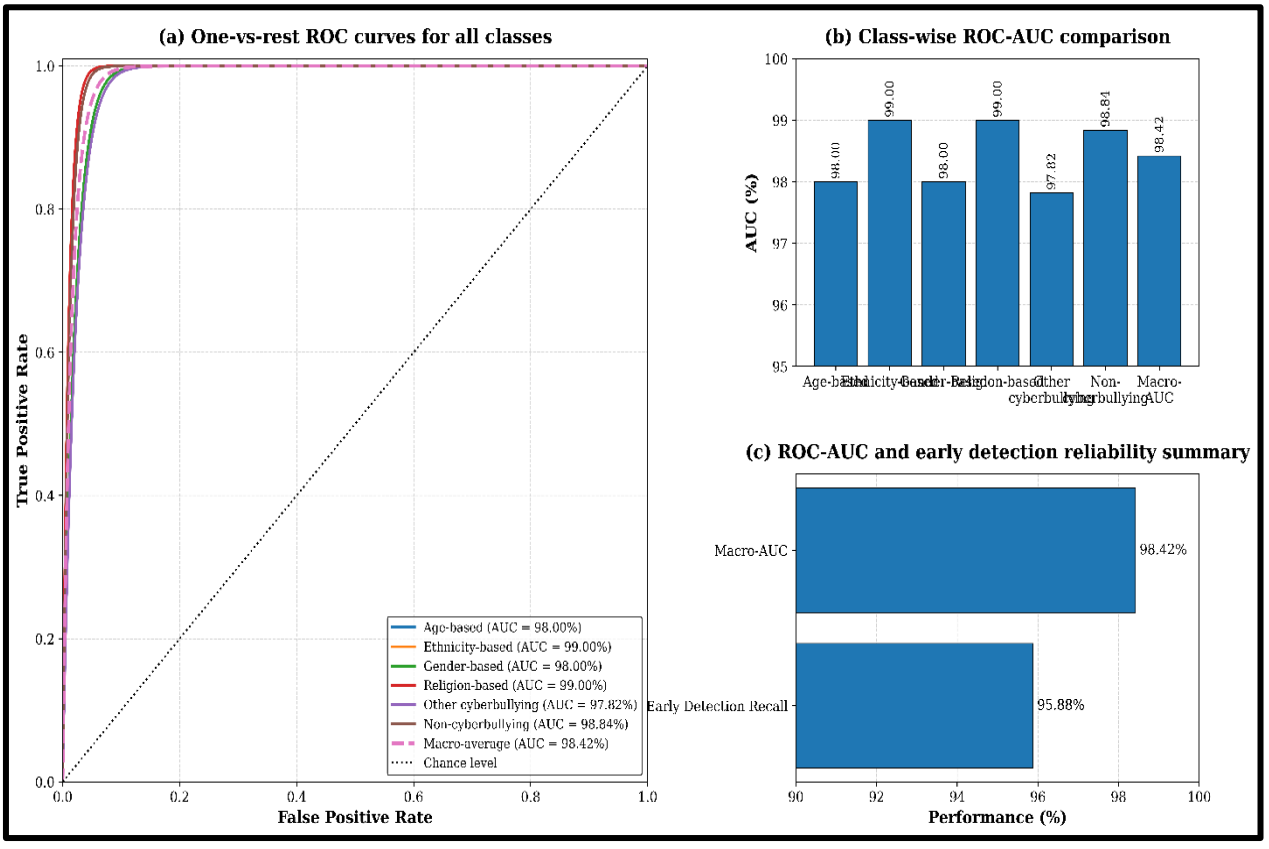

Figure 7. One-vs-rest ROC curves for multiclass cyberbullying classification.

The early detection recall of the proposed model was 95.88%, indicating that the model was able to detect most of the harmful posts based on the current text alone, without relying on the user history or future thread information. This is very useful as online abuse can rapidly ramp up after the release. A first-stage detection model can then help to speed up the moderation, reporting or prioritization of reviews in Figure 7.

4.5 Ablation Study

The ablation results in Table 12 give the contribution from each model component. The results showed that removing the auxiliary features lowered the macro-F1 to 96.05% and using auxiliary features without BERT gave a result of 89.48%

macro-F1. This is a confirmation that semantic representation is still the primary performance factor, but auxiliary features can make a significant contribution in conjunction with BERT embeddings.

Table 12. Ablation analysis of the proposed framework.

Variant	Configuration	Accuracy (%)	Macro-F1 (%)	Performance Observation
V1	TF-IDF + XGBoost	88.91	88.64	Limited semantic understanding
V2	BERT + Dense	94.31	94.08	Strong transformer baseline
V3	Auxiliary Features + XGBoost	89.84	89.48	Useful but insufficient alone
V4	BERT Embeddings + XGBoost	95.18	94.94	Better structured classification
V5	BERT + Auxiliary Features + XGBoost	96.18	96.05	Best balance of semantic and behavioural cues

The difference between V4 and V5 indicates that the improvement is not due to the change of dense classifier to XGBoost. The auxiliary feature stream provided valuable information about the style of bullying, emotional tone and social media expression. This aligns with the design decision in the proposed framework and with previous studies that showed sentiment, context, and explainability-related features are useful to enhance the detection of harmful content [16], [24], [33], [34].

4.6 Statistical Validation

Each model was tested with five independent random seeds to assess its stability of improvement. Selected metrics are reported in the mean, standard deviation and 95% confidence interval in Table 13. The proposed model had a low variance, suggesting consistent behaviour when the model is run multiple times.

Table 13. Statistical validation of the proposed model over five random seeds.

Metric	Mean (%)	Std. Dev.	95% Confidence Interval
Accuracy	96	0.21	95.99–96.37
Macro-F1	96	0.24	95.84–96.26
Weighted-F1	96	0.2	95.98–96.34
Early Detection Recall	96	0.27	95.64–96.12
Macro-AUC	98.42	0.16	98.28–98.56

The paired significance test between the proposed model and the best baseline model (BERT Embeddings + XGBoost) yielded ($p<0.05$). This indicates that the performance improvement was not a random occurrence when initialising or partitioning data. The very tight confidence intervals also lend credibility to the framework proposed.

4.7 Explainability and Error Inspection

The explainability analysis revealed that BERT contextual features were the most significant features in the final prediction, while auxiliary features also had a noticeable impact on cases that were on the edge of the prediction. Profanity score, sentiment polarity, punctuation intensity, uppercase ratio, and mention count had the most impact on the auxiliary features.

The inspection of errors revealed three typical error patterns. Firstly, posts that were sarcastic but not containing any explicit abusive language were sometimes considered as non-cyberbullying. Secondly, short posts with minimal context led to confusion between “other cyberbullying” and identity-based categories. Third, code-mixed or misspelled insults made it difficult to understand the meaning of the words. Based on these results, further enhancements are needed such as modelling for conversations, multilingual normalization, and learning features that are sensitive to sarcasm. Figure 8 shows the grouped SHAP-based explanation of BERT contextual and auxiliary feature contributions.

ISSN: 3079-5354

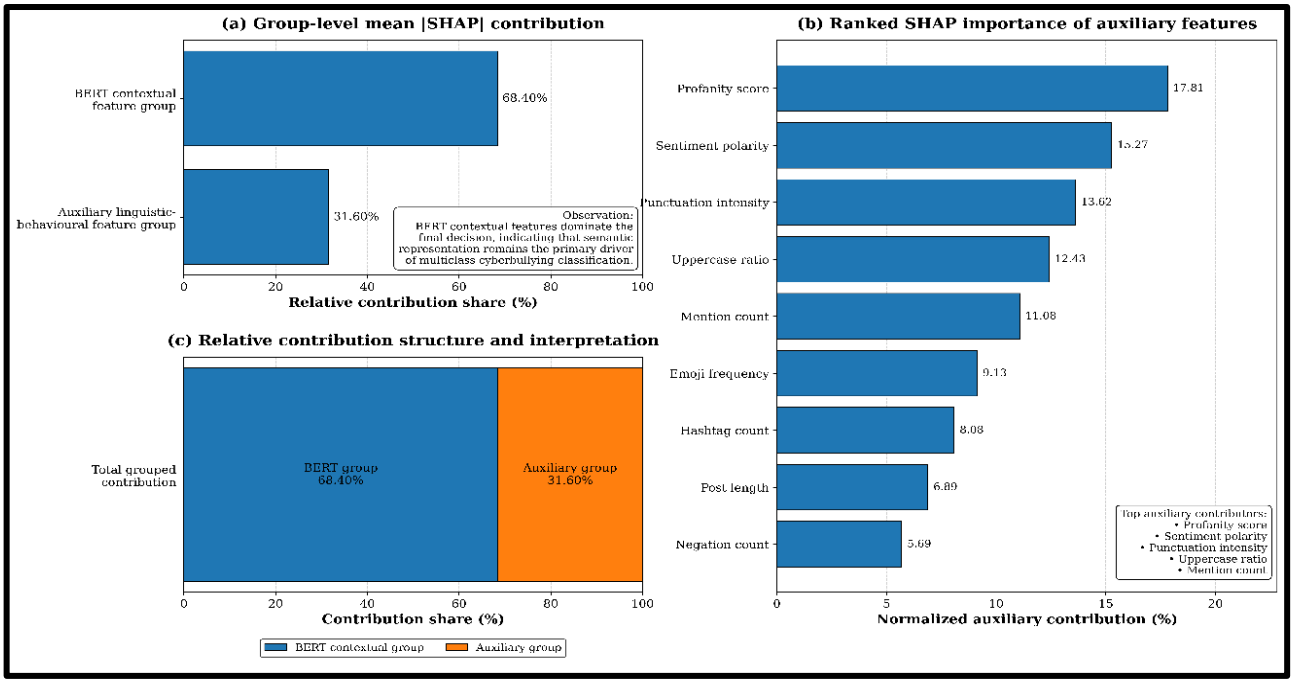

Figure 8. Grouped SHAP-based explanation of BERT contextual and auxiliary feature contributions.

4.8 Comparative Interpretation

The proposed framework was superior to the conventional machine-learning methods as it was not solely based on sparse lexical patterns. It enhanced the reliability of the classification at the category level with structured linguistic and behavioural signals when compared to transformer-only classification. This finding aligns with the trend of hybrid, context-aware and explainable cyberbullying detection systems [2], [5], [15], [16], and [19]. A cyberbullying model should be able to identify offensive text as well as be able to differentiate the type of harm with reasonable accuracy. The proposed Hybrid BERT–XGBoost framework was able to achieve this balance, through the integration of semantic understanding, auxiliary social media cues, and regularized multiclass decision learning. In general, the results show that the proposed model is appropriate to detect cyberbullying in short social media text, both in early and in category.

5. Discussion

The findings show that the proposed Hybrid BERT–XGBoost framework outperforms the baseline models, including the machine-learning and deep-learning models, and the transformer-only model, in terms of a more balanced and reliable early detection of cyberbullying. The improvement in macro-F1 and class-wise recall indicates that the model is not simply improving recognition of the overall class balance, but also the recognition of sensitive bullying types like religion-, ethnicity-, gender- and age-based bullying. This is significant as it is not only important to have high overall accuracy but also have reliability at the category level for practical moderation systems. The performance improvement seems to be due to the complementary representation. BERT understands the meaning in context and implicit semantic signals, while auxiliary linguistic and behavioural signals retain social-media-specific signals like sentiment polarity, mentions, emojis, capitalization, punctuation stress and profanity. XGBoost also learns nonlinear interactions among these features to further enhance the stability of the decisions. The ablation results show that the fusion of BERT embeddings and auxiliary features is the most robust, while neither of them is as robust individually.

5.1 Limitations

There are a few limitations to this study. First, the main data set consists of short social media text, so the model might not fully represent longer social media conversations or the social media interaction dynamics. Second, expressions of sarcastic, coded and very implicit cyberbullying are hard to accurately categorize. Third, while auxiliary features can help to make the embedding more interpretable, the dimensions of BERT embeddings are not human-readable. Lastly, the framework

was tested primarily using text data, and multimodal signals like images, videos, memes and users' interaction history were not considered. Future work should take into account these issues by validating the system in multiple languages, multiple media, and conversation-aware.

6. Conclusion and Future Scope

This study introduced a Hybrid BERT–XGBoost model to detect and classify cyberbullying in social media text in an early stage and multiclass classification. The proposed model integrates the contextual embedding from BERT with other linguistic and behavioural features, such as sentiment, profanity, punctuation intensity, capitalization, hashtags, mentions, emojis, and post length. The results indicate that the fusion has a better performance than the classical machine-learning, deep-learning, and transformer-only baselines in terms of category-level detection. Using XGBoost further enhances the stability of the classification and offers more opportunities for interpretation at the feature level. The study provides a practical and meaningful cyberbullying detection pipeline at the first post-observation level to identify various types of cyberbullying. There are some restrictions on sarcastic, code-mixed and context-dependent abuse, however. Future work will be based on the framework, which will be extended with the use of multilingual datasets, conversation-level context, multimodal cues and real-time platform validation. Other studies could also be conducted on lightweight deployment for scalable moderation systems.

Acknowledgements

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU263669).

Funding

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU263669).

Contributions

All authors have equal contribution in this manuscript

Ethics declarations

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests.

References

[1] Hyder, S. B., Tariq, N., Moqurrab, S. A., Ashraf, M., Yoo, J., & Srivastava, G. (2024). BERT-based deceptive review detection in social media: Introducing DeceptiveBERT. *IEEE Transactions on Computational Social Systems, 11*(6), 7234–7243. https://doi.org/10.1109/TCSS.2024.3403937

[2] Abdullah Alotaibi, E., & Al-Samawi, A. (2025). Cyberbullying detection and identification using machine learning-based hybrid framework. *IEEE Access, 13*, 215423–215437. https://doi.org/10.1109/ACCESS.2025.3634347

[3] Razi, F., & Ejaz, N. (2024). Multilingual detection of cyberbullying in mixed Urdu, Roman Urdu, and English social media conversations. *IEEE Access, 12*, 105201–105210. https://doi.org/10.1109/ACCESS.2024.3432908

[4] Mamodiya, U., Kishor, I., Naz, R., Almaiah, M., & Alqutaish, A. (2026). A hybrid blockchain-based framework for adaptive cyber-risk prediction and multi-layer threat mitigation in enterprise networks. *Journal of Cybersecurity and Privacy, 6*(3), Article 85. https://doi.org/10.3390/jcp6030085

[5] Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access, 11*, 55533–55560. https://doi.org/10.1109/ACCESS.2023.3275130

[6] Mamodiya, U., Kishor, I., Garine, R., Ganguly, P., & Naik, N. (2025). Artificial intelligence based hybrid solar energy systems with smart materials and adaptive photovoltaics for sustainable power generation. *Scientific Reports, 15*(1), Article 17370. https://doi.org/10.1038/s41598-025-01788-4
[7] Abusaqer, M., Saquer, J., & Ghosh, M. (2026). BERT-OTA: Enhancing hate speech detection with ontology-guided transformer attention. *IEEE Access, 14*, 3345–3358. https://doi.org/10.1109/ACCESS.2026.3650874
[8] Kishor, I., & Syed, A. A. (2026). A novel federated architecture integrating ViT and BioBERT for real-time healthcare diagnosis. In *Transformative role of transformer models in healthcare* (Chap. 1, pp. 1–24). IGI Global. https://doi.org/10.4018/979-8-3373-2038-0.ch001
[9] Yadav, A., & Singh, V. (2025). HateFusion: Harnessing attention-based techniques for enhanced filtering and detection of implicit hate speech. *IEEE Transactions on Computational Social Systems, 12*(4), 1700–1715. https://doi.org/10.1109/TCSS.2024.3512573
[10] Kishor, I., Mamodiya, U., Almaayah, M., Alqutaish, A., Shehab, R., & Obeidat, M. (2025). Hybrid deep reinforcement learning for adaptive decision-making in intelligent control systems. *Engineered Science, 38*, Article 1680. https://doi.org/10.30919/es1680
[11] Alfurayj, H. S., Lebai Lutfi, S., & Perumal, R. (2024). A chained deep learning model for fine-grained cyberbullying detection with bystander dynamics. *IEEE Access, 12*, 105588–105604. https://doi.org/10.1109/ACCESS.2024.3435840
[12] Ismail, W. S., Ullah, H., Adnan, M., & Ullah, F. (2026). Multilingual multimodal cyberbullying detection through adaptive and hierarchical fusion. *Array, 29*, Article 100689. https://doi.org/10.1016/j.array.2026.100689
[13] Ejaz, N., Razi, F., & Choudhury, S. (2024). Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm. *Computers in Human Behavior, 153*, Article 108123. https://doi.org/10.1016/j.chb.2023.108123
[14] Mamodiya, U., Redkar, S., Sharma, R., Kishor, I., & Goyal, P. (2025). A hybrid machine learning framework for predictive maintenance in renewable energy systems to enhance efficiency and longevity. In *Proceedings of the 2025 International Conference on Sustainability, Innovation & Technology (ICSIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSIT65336.2025.11295132
[15] Yadavalli, U. S., & Sahoo, S. R. (2026). A multi-granular hybrid neural architecture for detecting abusive content in online social networks (OSNs) with contextual awareness. *Journal of Big Data, 13*(1), Article 5. https://doi.org/10.1186/s40537-025-01343-y
[16] Philipo, A., Ding, J., Sarwatt, D., Mohamed, J., Yusufu, A., Daneshmand, M., & Ning, H. (2026). Sentiment-enhanced cyberbullying detection models on social media platforms. *ACM Transactions on the Web, 20*(1), 1–26. https://doi.org/10.1145/3766075
[17] Mathur, V., Saini, Y., Giri, V., Choudhary, V., Bharadwaj, U., & Kumar, V. (2021). Weather station using Raspberry Pi. In *Proceedings of the 2021 Sixth International Conference on Image Information Processing (ICIIP)* (pp. 279–283). IEEE.
[18] Ashiq, W., Kanwal, S., Rafique, A., Waqas, M., Khurshaid, T., Montero, E. C., ... & Ashraf, I. (2024). Roman urdu hate speech detection using hybrid machine learning models and hyperparameter optimization. *Scientific Reports, 14*(1), 28590. https://doi.org/10.1038/s41598-024-79106-7
[19] Mahajan, E., Mahajan, H., & Kumar, S. (2024). EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media. *Expert Systems with Applications, 236*, Article 121228. https://doi.org/10.1016/j.eswa.2023.121228
[20] Mamodiya, U., Kishor, I., Guler, N., Hindi, J., & Naik, N. (2025). Implementation of reinforcement learning environment for hybrid renewable energy systems. In *Proceedings of the 2025 International Conference on Computational Intelligence, Security, and Artificial Intelligence (IntelliSecAI)* (pp. 1–6). IEEE. https://doi.org/10.1109/IntelliSecAI66368.2025.11472894
[21] Putra, D., & Wang, H.-C. (2024). Semi-meta-supervised hate speech detection. *Knowledge-Based Systems, 287*, Article 111386. https://doi.org/10.1016/j.knosys.2024.111386
[22] Talaat, S. (2023). Sentiment analysis classification system using hybrid BERT models. *Journal of Big Data, 10*, Article 110. https://doi.org/10.1186/s40537-023-00781-w
[23] Kishor, I., Mamodiya, U., Sharma, M., Kumar, G., & Goyal, P. (2025). Integrating blockchain with IoT for secure and transparent supply chain management in sustainable manufacturing. In *Proceedings of the 2025 International Conference on Sustainability, Innovation & Technology (ICSIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICSIT65336.2025.11294611
[24] Kibriya, H., Siddiqa, A., Khan, W. Z., & Khan, M. K. (2024). Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. *Computers & Electrical Engineering, 116*, Article 109153. https://doi.org/10.1016/j.compeleceng.2024.109153
[25] Lerotholi, A., & Obagbuwa, I. C. (2025). Sentiment analysis to detect cyberbullying on Twitter. *Human Behavior and Emerging Technologies, 2025*, Article 5419912. https://doi.org/10.1155/hbe2/5419912
[26] Ahuja, V., Kishor, I., Alqutaish, A., Shehab, R., & Obeidat, M. (2026). Mitigating information leakage risks in secure multiparty computation through function hiding. *Journal of Cyber Security and Risk Auditing, 2026*(1), 38–72. https://doi.org/10.63180/jcsra.thestap.2026.1.3
[27] Miao, Z., Chen, X., Wang, H., Tang, R., Yang, Z., Huang, T., & Tang, W. (2023). Detecting offensive language based on graph attention networks and fusion features. *IEEE Transactions on Computational Social Systems*, *11*(1), 1493-1505. https://doi.org/10.1109/TCSS.2023.3250502
[28] Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., ... & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, *14*(1), 204. https://doi.org/10.1007/s13278-024-01361-3
[29] Ortiz Salazar, A. (2025). Detecting hate crimes through machine learning and natural language processing. *Police Practice and Research, 26*(6), 746–768. https://doi.org/10.1080/15614263.2024.2397363
[30] Mishra, L., Sinha, S., & George, C. P. (2024). Shielding against online harm: A survey on text analysis to prevent cyberbullying. *Engineering Applications of Artificial Intelligence, 133*, Article 108241. https://doi.org/10.1016/j.engappai.2024.108241

[31] Umansky, N., Kubli, M., Kotarcic, A., Bronner, L., Kurer, S., Grech, P., ... & Donnay, K. (2026). Improving hate speech detection with large language models. *European Journal of Political Research*, *1*, 12. https://doi.org/10.1017/S1475676525100546
[32] Zou, L., He, Z., Zhou, C., & Zhu, W. (2024). Multi-class multi-label classification of social media texts for typhoon damage assessment: A two-stage model fully integrating the outputs of the hidden layers of BERT. *International Journal of Digital Earth*, *17*(1). https://doi.org/10.1080/17538947.2024.2348668
[33] Maity, K., Jain, R., Jha, P., & Saha, S. (2024). Explainable cyberbullying detection in Hinglish: A generative approach. *IEEE Transactions on Computational Social Systems*, *11*(3), 3338–3347. https://doi.org/10.1109/TCSS.2023.3333675
[34] Hussain, I., Rizvi, M. R., Abbas, Z., Cheema, A. N., & Almanjahie, I. M. (2025). MUST: An explainable AI-based framework for multilingual hate speech detection. *IEEE Access*, *13*, 202758–202778. https://doi.org/10.1109/ACCESS.2025.3629527
[35] Niu, Y., Chen, S., Kökciyan, N., & Qiu, W. (2025). Analyzing social media comments to understand and detect privacy violations. *IEEE Transactions on Computational Social Systems*, *12*(5), 2661–2674. https://doi.org/10.1109/TCSS.2024.3521936
[36] Ghosh, S., Priyankar, A., Ekbal, A., & Bhattacharyya, P. (2023). A transformer-based multi-task framework for joint detection of aggression and hate on social media data. *Natural Language Engineering*, *29*(6), 1495–1515. https://doi.org/10.1017/S1351324923000104
[37] Yu, K., Zhu, X., Guo, Z., Tolba, A., Rodrigues, J. J. P. C., & Leung, V. C. M. (2024). A cross-field deep learning-based fuzzy spamming detection approach via collaboration of behavior modeling and sentiment analysis. *IEEE Transactions on Fuzzy Systems*, *32*(12), 7168–7182. https://doi.org/10.1109/TFUZZ.2024.3425510
[38] Sánchez-Corcuera, R., Zubiaga, A., & Almeida, A. (2024). Early detection and prevention of malicious user behavior on Twitter using deep learning techniques. *IEEE Transactions on Computational Social Systems*, *11*(5), 6649–6661. https://doi.org/10.1109/TCSS.2024.3419171
[39] Vujičić Stanković, S., & Mladenović, M. (2023). An approach to automatic classification of hate speech in sports domain on social media. *Journal of Big Data*, *10*, Article 109. https://doi.org/10.1186/s40537-023-00766-9
[40] Hashmi, E., Yayilgan, S. Y., Yamin, M. M., & Ullah, M. (2025). Enhancing misogyny detection in bilingual texts using explainable ai and multilingual fine-tuned transformers. *Complex & Intelligent Systems*, *11*(1), 39. https://doi.org/10.1007/s40747-024-01655-1
[41] Saleous, H., Gergely, M., & Shuaib, K. (2025). Exploring NLP-based solutions to social media moderation challenges. *Human Behavior and Emerging Technologies*, *2025*, Article 9436490. https://doi.org/10.1155/hbe2/9436490
[42] Al-Hashedi, M., Soon, L.-K., Goh, H.-N., Lim, A. H. L., & Siew, E.-G. (2023). Cyberbullying detection based on emotion. *IEEE Access*, *11*, 53907–53918. https://doi.org/10.1109/ACCESS.2023.3280556
[43] Alfurayj, H. S., Farid, D. M., Luna-Jiménez, C., & Lebai Lutfi, S. (2026). CYBY24 and step-wise model for thread-based fine-grained cyberbullying detection. *IEEE Access*, *14*, 10351–10370. https://doi.org/10.1109/ACCESS.2026.3652469
[44] Kapil, P., Kumari, G., Ekbal, A., Pal, S., Chatterjee, A., & Vinutha, B. N. (2023). HHSD: Hindi hate speech detection leveraging multi-task learning. *IEEE Access*, *11*, 101460–101473. https://doi.org/10.1109/ACCESS.2023.3312993
[45] Sweidan, S., Farouk, N. A., Abouhawwash, M., Askar, S. S., & Taha, M.(2026). DeBERTa-based framework for detecting machine-generated content on social media: A comparative study. Journal of Big Data, 13, 10. https://doi.org/10.1186/s40537-025-01349-6