IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Multimodal Cross-Attention Fusion Model for Explainable Cyberbullying Identification on social media

Ms. Srishti Tripathi [1]
Department of Computer Science and Engineering, Oriental Institute of Technology, Bhopal, Madhya Pradesh, India

Dr. Rajesh K. Shukla [2]
Department of Computer Science and Engineering, Oriental Institute of Technology, Bhopal, Madhya Pradesh, India

Mr. Shivank Kumar Soni [3]
Department of Computer Science and Engineering, Oriental Institute of Technology, Bhopal, Madhya Pradesh, India

Abstract— The variety and context-dependent nature of abusive information on social media has made cyberbullying a serious problem. Using transformer-based feature extraction, cross-modal attention, and gated feature fusion, this research presents a cross-modal fusion framework for multi-class cyberbullying detection that combines verbal and visual representations. The textual content is encoded using XLM-RoBERTa-base, while visual information is represented using CLIP ViT-B/32. The collected features are projected into a shared 512-dimensional fusion space, where responses between textual and visual representations are captured via bidirectional multi-head cross-attention with eight attention heads. The modality-specific representations are adaptively combined by a learnable gated fusion method before being classified into five categories.. On the validation set of 6,113 posts, the proposed model achieves 91.61% accuracy, 91.92% macro-precision, 91.46% macro-recall, and 91.53% macro-F1, slightly outperforming the DistilBERT baseline, which achieves 91.59% accuracy and 91.49% macro-F1. Class-wise evaluation shows strong performance for age (97.92% F1) and ethnicity (97.22% F1), while SHAP and LIME analyses provide interpretability by identifying influential textual features. These results demonstrate the effectiveness of suggested fusion system for multi-class cyberbullying identification and provide a foundation for multimodal social-media content analysis.

Keywords— Cyberbullying Detection; Multimodal Learning; XLM-RoBERTa; CLIP; Cross-Modal Attention; Gated Feature Fusion.

I. INTRODUCTION

In the present world, social media (SM) plays a crucial role in definition and dissemination of information and material. Additionally, social media platforms facilitate virtual communities that are built on relationships and interests, allowing users to interact and network. It is impossible to overstate significance of social media's recent rise to prominence as a platform for connection and communication. These social media sites' growing popularity (Facebook, Twitter, Instagram, Tinder, etc.) makes it possible for users to share a variety of multimedia messages. E.g., activities of these SM platforms were crucial during COVID-19 epidemic because individuals in different areas could quickly share real-time information about occurrences and findings [1].

With billions of users worldwide, social media platforms have become essential to contemporary communication. However, the development of abusive online behavior known as cyberbullying—the persistent use of digital communication to harass, denigrate, or threaten people or groups—

has also been made possible by this unprecedented connectedness [2]. Cyberbullying is linked to low self-esteem, anxiety, sadness, and in extreme situations, self-harm and suicide thoughts among victims, according to earlier surveys, highlighting the need for prompt and accurate detection mechanisms [3]. Cyberbullying is often expressed through implicit, sarcastic, context-dependent, and category-specific language, making reliable detection more difficult than simple keyword-based identification[4] [5][6].

Traditional moderation mechanisms such as user-driven blocking and reporting are inherently reactive and struggle to keep pace with the volume and velocity of user-generated content. This has motivated a large body of research into automated cyberbullying identification using Natural Language Processing (NLP)[7] and Machine Learning (ML)[8], progressing from keyword filters and classical classifiers toward fine-tuned transformer-based Large Language Models (LLMs) e.g. BERT and its variants [9]. Nevertheless, most existing approaches operate on text alone, are evaluated as coarse binary (bullying / non-bullying) classifiers, and provide little to no explanation of their predictions, which limits their trustworthiness and adoption by platform moderators and policy makers.

Beyond individual harm, cyberbullying erodes trust in online communities and can trigger repetitive, escalating abusive behaviour when left unmoderated, with recent industry reports pointing to sustained growth in harassment complaints across major platforms[10]. Existing platform-side mechanisms: flagging, blocking, and manual reporting — are inherently reactive and place the burden of detection on victims, which further motivates research into proactive, automated, and ideally explainable detection systems that can support, rather than replace, human moderators[11]. Beyond binary flagging, distinguishing the specific target category of abuse (e.g., gender-, religion-, age-, or ethnicity-based) is itself valuable for moderation teams, as it enables targeted intervention policies and more informative reporting to affected communities, a capability largely absents from prior binary-classification systems[12]. This paper addresses these gaps by proposing a multimodal cross-attention fusion architecture that combines a pretrained multilingual text encoder with a pretrained vision-language encoder enabling the system to be extended to posts that combine text and imagery while, on the present text-only benchmark, still exploiting rich contextual representations that text-image pretraining affords.

A. Motivation and Contribution

Cyberbullying has become more widespread due to quick expansion of SM platforms, making automatic and trustworthy detection more crucial. Existing approaches often rely mainly on textual information and may struggle with the semantic ambiguity of different cyberbullying categories, particularly when distinguishing harmful content from non-cyberbullying posts. Moreover, conventional multimodal methods may not effectively model the interaction between textual and visual information. Therefore, this work is motivated by the need for a robust framework that can capture contextual textual features and provide flexible mechanism to multimodal feature interaction and adaptive fusion.

The following research contribution of this work are:

- A transformer-based cross-modal fusion framework is proposed for five-class cyberbullying detection using XLM-RoBERTa and CLIP representations.
- To discover how textual and visual representations interact in a shared 512-dimensional environment, a bidirectional cross-modal attention mechanism (BCAM) with eight attention heads is presented.
- A learnable gated fusion mechanism is developed to adaptively control the contribution of modality-specific features before classification.
- The proposed model is comprehensively evaluated against a DistilBERT baseline using accuracy, ROC-AUC, confusion matrices, and class-wise analysis.
- SHAP and LIME-based explainability analyses are incorporated to identify influential textual features and provide insights into model's classification decisions.

B. Organization of paper

The rest of this paper is prepared as occurs. Section I introduces research problem, motivation, and contributions. Section II presents related work on cyberbullying and multimodal detection. Section III defines the proposed procedure and investigational setup. Section IV presents results, and explainability findings. Section V accomplishes paper and outlines future work.

II. LITERATURE REVIEW

Recent cyberbullying research has increasingly moved from predictable ML approaches toward hybrid deep-learning and transformer-based architectures. Cao et al. (2026) proposed a BERT-BiGRU-CNN (BBGC) model that combines BERT contextual embeddings, BiGRU sequential

features, and CNN-based local features, with the fused representation used for final cyberbullying classification[13]. Pal, Kondekar, and Khan (2026) combined TF-IDF, word embeddings, ensemble classifiers, and BERT, reporting 94.00% accuracy and BERT-based F1-scores above 0.92[14]. Kumari and Kaur (2025) achieved 98.19% baseline accuracy with 96.37% with federated learning as well as differential privacy using a privacy-preserving system that included LIME, federated learning, transfer learning, with differential privacy [15]. Garc and Arriba-p (2024) explored stream-based machine learning with LLM-based feature engineering and explainability, obtaining performance close to 90% across evaluation metrics[16]. Similarly, Muneer et al. (2023) proposed a modified BERT model, achieving 97.4% accuracy, 0.964 F1-score, 0.950 precision, and 0.92 recall on Twitter data, while performance reached 90.97% on combined Twitter and Facebook data[17].

Other studies have addressed specific linguistic and feature-engineering challenges. Dewani et al. (2023) investigated Roman Urdu cyberbullying using statistical features, word and combined n-grams, and TF-IDF-based representations; SVM with hybrid n-gram features achieved approximately 83% accuracy[18]. Basheer (2022) incorporated semantic, sentiment, syntactic, pragmatic, TF-IDF, and CountVectorizer features with classifiers for example MLP, LR, NB, KNN, RF, and SVM[19]. Alotaibi et al. (2021) proposed a multichannel architecture combining BiGRU, Transformer, and CNN for aggressive-content detection and reported approximately 88% accuracy[20].

Research gaps: Although these approaches demonstrate strong results, most concentrate on textual information or single-modal representations. Limited work explicitly models the interaction between text and visual content over bidirectional cross-modal attention and adaptive gated fusion. Moreover, class-level ambiguity, particularly between non-cyberbullying and targeted cyberbullying categories, remains insufficiently addressed. These limitations motivate the proposed XLM-RoBERTa–CLIP framework, which integrates contextual textual and visual representations through attention and learnable gated fusion.

III. METHODOLOGY

The proposed framework (figure 1) performs cyberbullying detection through a pipeline consisting of data preprocessing, multimodal representation learning, feature fusion, classification, and evaluation. The input tweets are cleaned by removing URLs, usernames, hashtags, punctuation, stopwords, emojis, and duplicate or invalid samples, followed by lemmatization and label encoding. The preprocessed text is represented using XLM-RoBERTa, while visual information is encoded using the CLIP ViT-B/32 encoder. The resulting text and visual features are projected into a common feature space and integrated using bidirectional multi-head cross-modal attention and a learnable gating mechanism. The fused representation is passed through fully connected layers with normalization and dropout to classify the input into five cyberbullying categories.. Performance is assessed using accuracy, confusion matrix, ROC, and precision–recall curves, while LIME, SHAP, attention weights, and gating weights are used to provide interpretability of model predictions.

A. Dataset Collection and Analysis

Experiments use public Kaggle Cyberbullying Classification dataset[1], comprising 47,692 tweets originally labelled into 6 categories. After removing 1,675 exact duplicate tweets and discarding ambiguous other_cyberbullying category, 5 classes remain: not_cyberbullying, gender, religion, age, and ethnicity, each represented by approximately 7,900–8,000 posts (Table I). Tweets shorter than four words or longer than 99 words were additionally removed to reduce noise from near-empty or anomalous posts, yielding a final corpus of 38,202 samples.

[1] https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

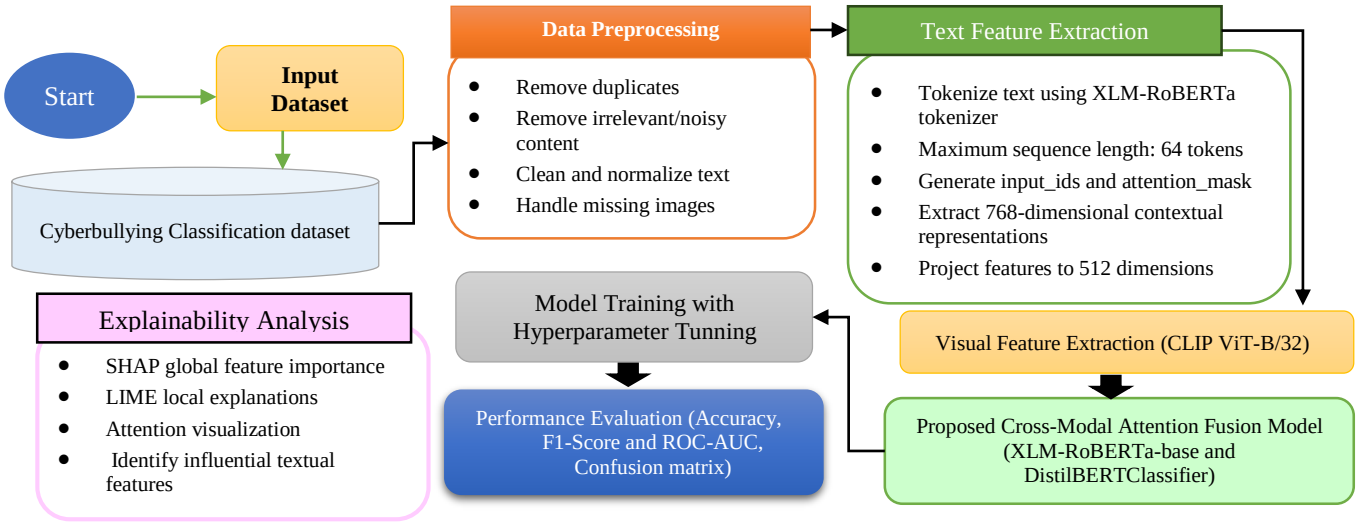

Fig. 1. Block Diagram of proposed work for cross-modal attention fusion pipeline for cyberbullying detection systems

TABLE I. CLASS DISTRIBUTION OF THE DATASET

Category	Raw Count	Share
Religion	7,995	16.8%
Age	7,992	16.8%
Ethnicity	7,952	16.7%
Not Cyberbullying	7,937	16.6%
Gender	7,898	16.6%
Other Cyberbullying (removed)	6,243	13.1%
Total (Raw)	**47,692**	**100%**
Total after de-duplication, class removal, and length filtering	**38,202**	—

Exploratory analysis (Fig. 2–3) shows that word count and average word length vary systematically by category — religion- and age-related posts tend to be noticeably longer on average than generic (not_cyberbullying) posts. Fig. 4 depicts the joint distribution of word count and average word length, while Fig. 5 shows per-class word clouds highlighting the most salient, category-discriminative vocabulary (e.g., slurs and identity terms concentrate in the gender, religion, and ethnicity classes), and Fig. 6 lists the overall ten most frequent terms in the corpus.

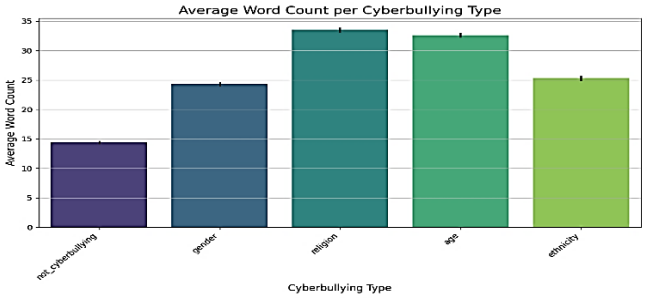

Fig. 2. Average word count per cyberbullying category.

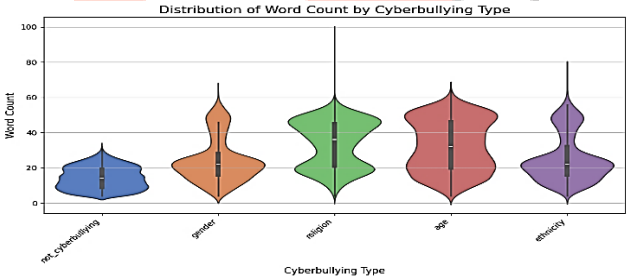

Fig. 3. Distribution of word count by cyberbullying category.

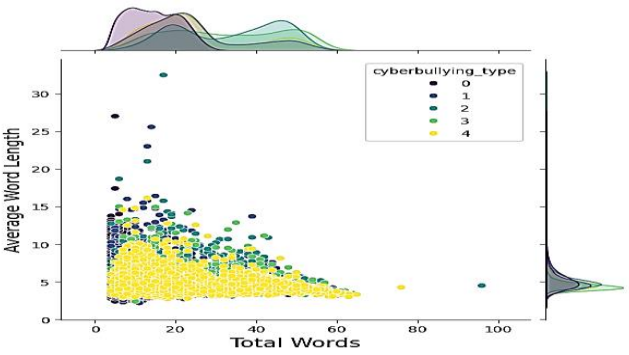

Fig. 4. Joint distribution of word count and average word length by category.

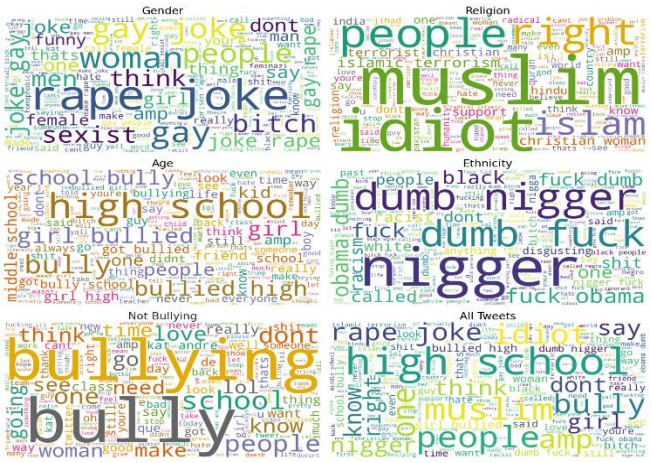

Fig. 5. Word clouds for the five cyberbullying categories.

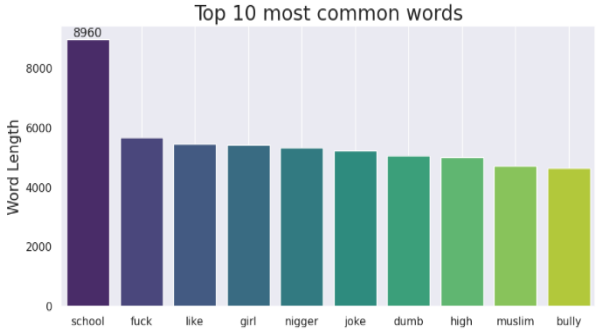

Fig. 6. Top-10 most frequent words in the corpus.

B. Data Preprocessing

The preprocessing stage removes duplicate tweets to avoid repeated information during model training; the code identifies 1,675 duplicate tweet entries and removes them using `drop_duplicates('tweet_text')`. The class distribution is then examined before model development. The final classification architecture considers five target categories—`not_cyberbullying`,`gender`, `religion`,`age`, and`ethnicity`—while the `other_cyberbullying` category is not included in the five-class model.

C. Text Tokenization and Feature Representation

The text processing module utilizes the pretrained XLM-RoBERTa-base transformer to extract contextual semantic features from the preprocessed tweets. Each tweet is first converted into subword tokens using the XLM-RoBERTa AutoTokenizer. To maintain a fixed input dimension, sequences are padded or truncated for max length of 64 tokens. The tokenizer generates two principal features: input_ids, which represent the numerical token indices, and attention_mask, which identifies valid tokens from padding positions. The tokenized input is then processed by the XLM-RoBERTa encoder to obtain contextualized token representations. Since XLM-RoBERTa-base produces a hidden representation of 768 dimensions for each token, the resulting feature matrix is represented as Eq.1:

$$H = \mathrm{XLM} - \mathrm{RoBERTa}(X, M), \quad H \in R^{T\times 768} \qquad (1)$$

where X denotes the token IDs, M denotes the attention mask, and H represents the contextual token-level feature matrix. Each 768-dimensional contextual feature is subsequently transformed into the common 512-dimensional fusion space using a learnable linear projection layer. This projection reduces the feature dimensionality and makes the textual representation compatible with the visual representation in the subsequent cross-modal fusion stage. The projected textual feature is calculated as Eq.2:

$$H_t = HW_t + b_t, \quad H \in R^{T\times 512} \qquad (2)$$

where $W_t \in R^{768\times 512}$signifies trainable projection matrix and $b_t \in R^{T\times 512}$ signifies bias vector. Thus, 768-dimensional contextual representation generated by XLM-RoBERTa is transformed into a 512-dimensional semantic feature representation, which is subsequently supplied to the cross-modal attention and gated fusion modules.

D. Visual Feature Extraction

The visual branch uses pretrained CLIP ViT-B/32 vision encoder (openai/clip-vit-base-patch32) to obtain image-level/patch-level representations. The CLIP vision encoder produces features with a hidden dimension of 768, which are transformed using a linear projection layer into the same 512-dimensional fusion space used by the textual branch. However, because the current Cyberbullying Tweets dataset does not contain an image column, the implementation sets image_col=None and generates a zero-valued tensor of size 3 × 224 × 224 when an image is unavailable. Therefore, the current implementation demonstrates the multimodal architecture using text-only data rather than providing genuine image information.

E. Cross-Modal Feature Fusion

The proposed model integrates the textual and visual representations via a BCMA. After projection, both modalities are represented in a common 512-dimensional feature space. Two multi-head attention operations are employed with **8 attention heads**, where the first attention layer enables textual features to attend to visual features, while second enables visual features to attend to textual features. For text-to-image direction, textual representation acts as query, while visual representation provides key and value. Conversely, for image-to-text attention, visual representation acts as the query and the textual representation

provides the key and value. The general scaled dot-product attention function is expressed as Eq.3:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where Q, K, and V represent query, key, and value matrices, correspondingly, and dk represents key dimensionality. Multi-head attention allows model to learn different relationships between the two modalities simultaneously. A dropout rate of **0.2** is used to attention mechanism. Following attention, residual connections combine original representation with the attended representation, and Layer Normalization is used to stabilize feature distributions. This bidirectional mechanism enables the model to capture complementary relationships among textual and visual data before relying on simple feature concatenation.

F. Gated Feature Fusion

Following cross-modal attention, the attended textual and visual representations are combined using a learnable gated fusion mechanism. The two 512-dimensional representations are concatenated to form a 1024-dimensional feature vector, which is passed through a linear transformation of 1024 → 512 followed by sigmoid activation. The resulting gate values lie between 0 and 1 and determine the contribution of each modality to the final representation. The gating function can be represented as Eq.4:

$$G = \sigma(W_g[H_t; H_v] + b_g) \quad (4)$$

where H_t and H_v denote the attended textual and visual features, [H_t; H_v v] represents their concatenation, W_g is learnable projection matrix, b_g is bias, and σ denotes sigmoid activation function. The gated representation is then generated by adaptively weighting the textual and visual information. This allows the network to emphasize the more informative modality while reducing the effect of less relevant or unavailable information.

G. Classification Function

The final fused representation has a dimensionality of **512** and is passed to a multilayer perceptron (MLP) classifier for cyberbullying category prediction. The classification network contains of FCLs with dimensions 512 → 256 → 128 → 64 → 5. ReLU activation functions are used in the hidden layers, while Batch Normalization is applied after the 512- and 256-dimensional layers to improve training stability. A dropout rate of 0.2 is also incorporated to reduce overfitting. The final fully connected layer produces five output logits, corresponding to five target classes. The predicted class is selected using the maximum output logit:

$$\hat{y} = \arg\max_{c\in\{1,\ldots..5,\}} z_c$$

where z_c represents output logit corresponding to class c, and y^ denotes final predicted cyberbullying category.

H. Proposed Cross-Modal Attention Fusion Model

The proposed classifier, CrossModalAttentionFusion, encodes text with XLM-RoBERTa-base (768-d hidden states) and images with the CLIP ViT-B/32 vision tower (768-d patch embeddings); when no image accompanies a post, a zero tensor of the shape expected by CLIP is substituted, allowing the same architecture to operate on text-only or text+image inputs without modification. Both modalities are linearly projected into a shared 512-dimensional fusion space. Two multi-head attention blocks (8 heads each) then let text tokens attend to image patches and vice versa, followed by layer normalisation and pooling. A learnable sigmoid gate, conditioned on the concatenated pooled text and image representations, weighs the relative contribution of each modality before fused vector is passed through four-layer MLP classifier (512→256→128→64 units, ReLU activations, batch normalisation, and 0.2 dropout at every stage) that outputs a probability distribution over five classes. The model additionally returns attention and gate weights that are later reused for explainability. In total fusion model has 433,173,510 parameters, all trainable end-to-end.

1) Baseline: DistilBERT Classifier

A text-only baseline, DistilBERTClassifier, extracts the [CLS] representation (768-d) from pretrained distilbert-base-uncased and feeds it through an MLP classification head that is architecturally identical to the fusion model's head (512→256→128→64 units, ReLU, batch normalisation, 0.2 dropout), ensuring that any performance difference is attributable to the encoder/fusion design rather than the classifier capacity. The baseline totals 66,930,949 parameters — approximately 6.5× fewer than the fusion model — providing a lightweight reference point.

I. Training Configuration

The filtered corpus (38,202 samples) is split with stratified sampling into 64% train, 16% val, and 20% test partitions (first an 80/20 train/test split, then an 80/20 split of training portion into train/val), using a fixed random seed of 42 throughout. For computational efficiency, model fitting uses a stratified demo subset of 3,000 train posts, while validation (6,113 posts) and evaluation are performed on the full corresponding partitions;

this trade-off is discussed as a limitation in Section V. Both models are optimised with AdamW (learning rate 2×10^{-5}), cross-entropy loss, batch size 16, and 20 training epochs, with the checkpoint of highest validation accuracy retained for final testing. Table II summarises all training hyperparameters.

TABLE II. CONFIGURATION AND PARAMETER SETTINGS OF THE PROPOSED FUSION MODEL AND DISTILBERT BASELINE

Parameter	Fusion Model	DistilBERT Baseline
Text Encoder	XLM-RoBERTa-base	DistilBERT-base-uncased
Image Encoder	CLIP ViT-B/32	—
Fusion Dimension	512	—
Cross-Attention Heads	8	—
Classifier Head	512-256-128-64 (MLP)	512-256-128-64 (MLP)
Dropout	0.2	0.2
Maximum_Sequence Length	64 tokens	64 tokens
Optimizer	AdamW	AdamW
LearningRate	2×10^{-5}	2×10^{-5}
LossFunction	Cross-Entropy	Cross-Entropy
BatchSize	16	16
Epochs	20	20
Train / Validation / Test Split	64% / 16% / 20% (Stratified)	64% / 16% / 20% (Stratified)
Training Subset Used	3,000 posts (Demo Mode)	24,450 posts (Full Split)
Random Seed	42	42
Total Parameters	433,173,510	66,930,949
Trainable Parameters	433,173,510	66,930,949

J. Model Performance Evaluation Measures

To assess the effectiveness of the suggested cyberbullying classification model based of performance measures. These metrics reflect overall prediction accuracy, class-wise identification capabilities, and discrimination performance. Precision shows how many samples projected as belonging to a specific class really belong to that class, whereas accuracy evaluates ratio of properly categorized samples to total number of samples. While F1-score offers a balanced metric between precision and recall, recall assesses the model's capacity to accurately identify real samples of a class. Macro-averaged Precision, Recall, and F1-score may be utilized to assign each of five cyberbullying classes in the proposed task equal weight. Furthermore, by analyzing the connection between the True Positive Rate (TPR) and False Positive Rate (FPR) at various decision thresholds, ROC-AUC (Receiver Operating Characteristic–Area Under the Curve) assesses the model's capacity to differentiate one class from the other classes. Equations 5 thru 9 define the related performance metrics:

$$\text{AccuracyACC)} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (5)$$

$$\text{Precision(PRE)} = \frac{\text{TP}}{\text{TP+FP}} \quad (6)$$

$$\text{Recall(REC)} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

$$\text{F1} - \text{score(F1)} = 2 * \frac{(\text{precision}*\text{recall}}{(\text{precision+recall})} \quad (8)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}, \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \quad (9)$$

ROC-AUC is calculated as area under ROC curve, where value closer to 1.0 indicates stronger class discrimination and a value near 0.5 indicates random discrimination. For five-class problem, ROC-AUC is evaluated using one-vs-rest strategy, with individual class AUC values aggregated using macro-averaging. Additionally, confusion matrix is used to identify class-wise correct and incorrect predictions, providing detailed insight into the classification behavior of the proposed model.

K. Evaluation and Explainability

To support model transparency, two complementary post-hoc explanation techniques are applied to both trained models. SHAP estimates each input token's contribution to the predicted class probability based on cooperative game theory, producing global feature-importance rankings as well as per-instance force plots. LIME perturbs input text and fits a local surrogate model to highlight words responsible for a specific prediction. Both techniques are applied to representative validation samples to qualitatively verify that the models attend to semantically relevant, category-discriminative vocabulary rather than spurious correlations.

L. Implementation Details

The full pipeline is executed in Python using PyTorch and the Hugging Face Transformers library for the XLM-RoBERTa, CLIP, and DistilBERT backbones; scikit-learn for label

encoding, data splitting, and evaluation metrics; and the shap and lime packages for explainability. Training was performed on a CUDA-enabled GPU runtime, with mixed-precision-free full-32-bit forward/backward passes for numerical stability across both models. All random operations (data splitting, demo sub-sampling, and explainer sampling) are seeded (seed = 42) to support reproducibility. The best checkpoint of each model, selected by validation accuracy, is persisted to disk (best_fusion_model.pt and best_distilbert_model.pt) and reloaded for the evaluation and explainability stages described in Section IV

IV. RESULTS ANALYSIS AND DISCUSSION

In this section provide results proposed system for cyberbullying classification. The proposed cross-modal fusion model and the DistilBERT baseline are assessed on held-out stratified validation set containing 6,113 posts, including 1,163 not_cyberbullying, 1,192 gender, 1,270 religions, 1,259 age, and 1,229 ethnicity samples. evaluation is performed using the classification report to provide both overall and class-wise performance.

A. Results of proposed models

The comparative results demonstrate that both transformer-based models achieve approximately 91.6% accuracy, confirming their effectiveness for cyberbullying detection. The proposed fusion model achieves 91.61% accuracy compared with 91.59% for DistilBERT and also provides higher macro-precision (91.92% vs. 91.71%) and macro-F1 (91.53% vs. 91.49%) (Table IV). The relatively small performance difference indicates that the textual information provides the majority of the discriminative capability for the present tweet-based dataset. Nevertheless, the proposed architecture provides an additional cross-modal attention and gated fusion mechanism, making it suitable for future datasets containing paired textual and visual information. The class-wise analysis further demonstrates that age and ethnicity are the most effectively classified categories, with F1-scores above 97%, whereas not_cyberbullying and gender remain comparatively challenging (Tables V and VI).

TABLE III. OVERALL CLASSIFICATION RESULTS ON THE VALIDATION SET (%)

Models	ACC	PRE (Macro)	REC (Macro)	F1 (Macro %)
Cross-modal Fusion (Proposed)	**91.61**	**91.92**	**91.46**	**91.53**
DistilBERT (Baseline)	91.59	91.71	91.44	91.49

TABLE IV. PER-CLASS CLASSIFICATION RESULTS OF PROPOSED FUSION MODEL

Class	PRE	REC	F1	Support
not_cyberbullying	75.78	87.70	81.31	1,163
gender	92.75	81.54	86.79	1,192
religion	94.77	94.09	94.43	1,270
age	**98.71**	**97.14**	**97.92**	1,259
ethnicity	97.62	96.83	97.22	1,229
Macro Average	**91.92**	**91.46**	**91.53**	**6,113**
Weighted Average	**92.15**	**91.61**	**91.72**	**6,113**

TABLE V. PER-CLASS CLASSIFICATION RESULTS OF THE DISTILBERT BASELINE

Class	Precision	Recall	F1-Score	Support
not_cyberbullying	77.39	86.24	81.57	1,163
gender	91.01	83.22	86.95	1,192
religion	94.75	93.78	94.27	1,270
age	97.01	97.78	97.41	1,259
ethnicity	**98.42**	96.18	**97.31**	1,229
Macro Average	**91.71**	**91.44**	**91.49**	**6,113**
Weighted Average	**91.87**	**91.59**	**91.66**	**6,113**

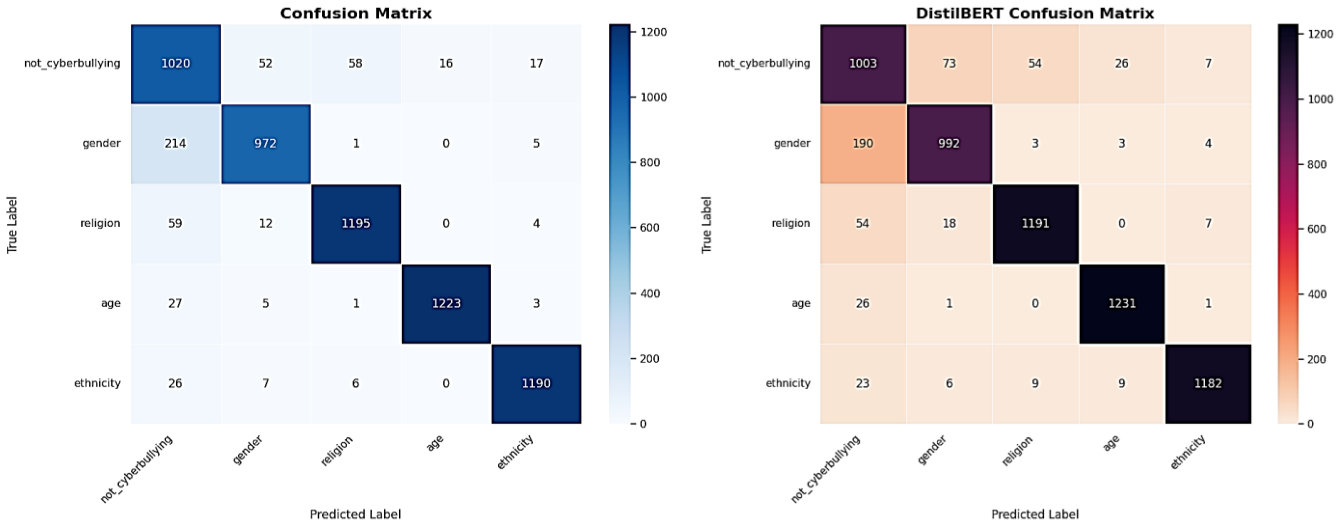

Fig. 7. Confusion matrix — fusion model and DistilBERT baseline (validation set).

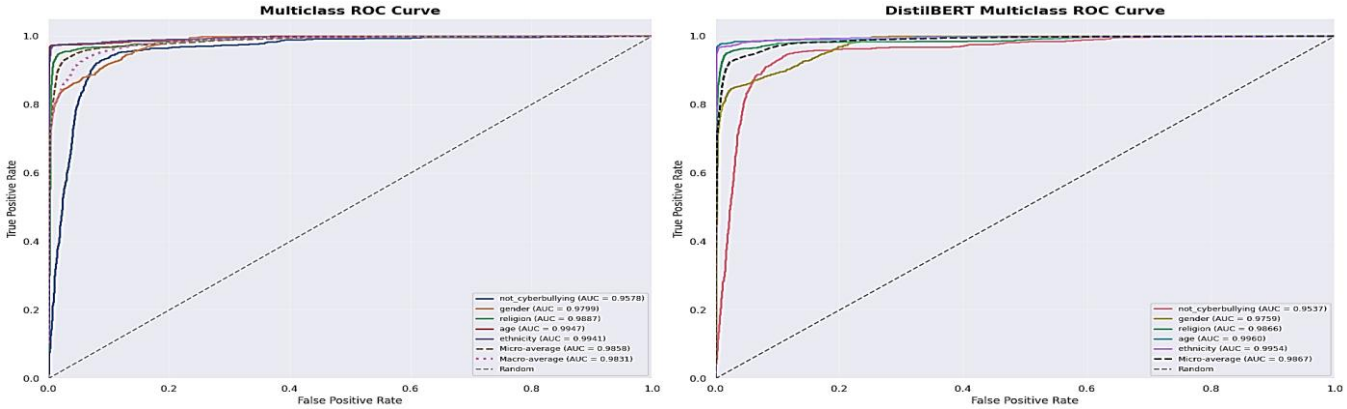

Fig. 8. Multiclass ROC curve — fusion model and DistilBERT baseline (validation set).

The confusion matrices in Fig. 7 show that major error occurs between gender and not_cyberbullying, with 214 gender samples misclassified as not_cyberbullying by the fusion model and 190 by DistilBERT, while age, religion, and ethnicity are comparatively well classified. The ROC curves in Fig. 8 further show that not_cyberbullying is the most difficult class to distinguish, with AUC values of 0.9578 for the fusion model and 0.9537 for DistilBERT, whereas age and ethnicity achieve AUC values above 0.99. Overall, the fusion model obtains 0.9858 micro-AUC and 0.9831 macro-AUC, while DistilBERT achieves a 0.9867 micro-AUC, indicating comparable discriminative performance between the two architectures.

B. Explainability Analysis

Fig. 9 report SHAP global feature-importance rankings (mean absolute SHAP value across sampled positions) for the text encoders of the fusion and DistilBERT models, respectively; in both cases, the earliest tokens of the input sequence carry disproportionate importance, consistent with the [CLS]/beginning-of-sequence pooling strategy used by both encoders. Fig. 10 show LIME explanations for a representative age-bullying example (“freshman year high school cheerleader bullied girl team thought lying mom cancer”): both models correctly predict the age class with near-certainty (fusion: 0.99; DistilBERT: 0.97 base value context), attributing the decision chiefly to the tokens bullied, high, and school, which aligns with human intuition about school-context bullying narratives. Attention visualisation of the fusion model's text-to-image cross-attention (not shown for space) confirms a near-uniform distribution over image patches on text-only inputs, as expected given the zero-filled image tensor, validating that the model correctly falls back to the textual channel when no visual evidence is present

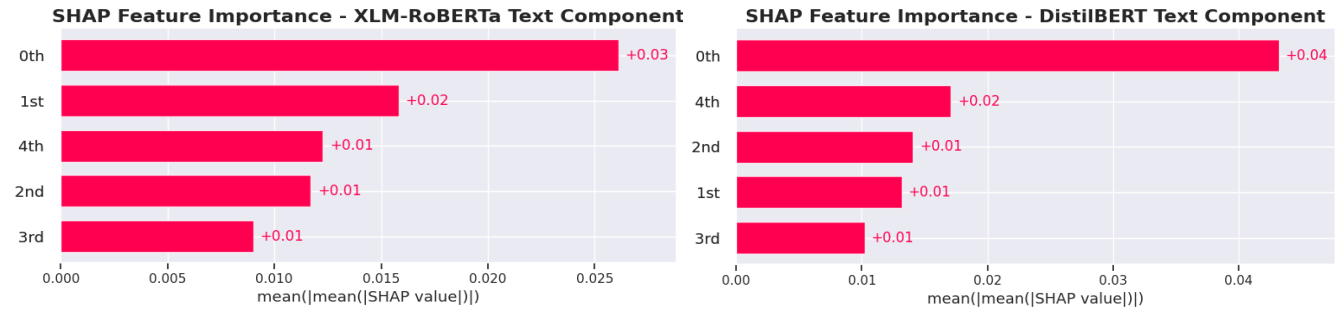

Fig. 9. SHAP feature importance — fusion model and DistilBERT baseline text component.

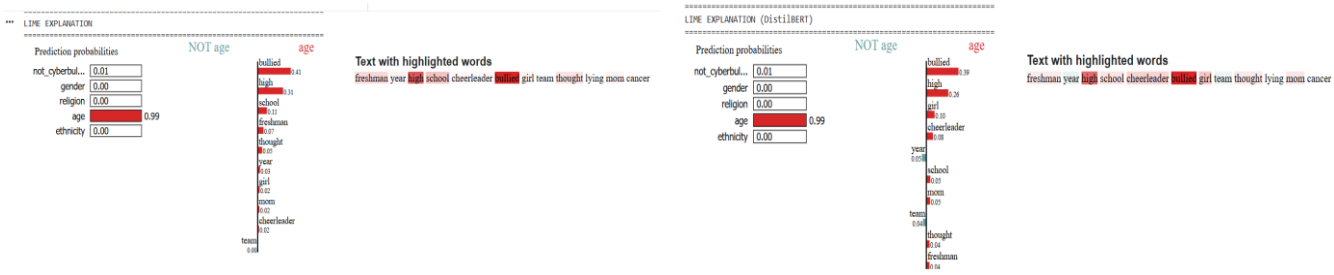

Fig. 10. LIME explanation for an age-related bullying example — fusion model and DistilBERT baseline text component.

C. Comparative Analysis

The outcomes in Table VII show that Cross-modal Fusion model gets best overall performance, with 91.61% ACC, 91.92% PRE, 91.46% REC, and 91.53% F1-score. It outperforms GNB, HATC, and ARFC in all major metrics. The DistilBERT baseline also performs competitively with 91.59% ACC and 91.49% F1-score, but remains slightly below the proposed model. These outcomes show performance of suggested method for cyberbullying detection.

TABLE VI. COMPARATIVE ANALYSIS OF PROPOSED MODEL WITH EXISTING APPROACHES

Works	Model	ACC	PRE	REC	F1-Score
S. Garc et al. [16]	GNB	84.46	85.18	84.52	84.40
	HATC	85.77	86.06	85.73	85.73
	ARFC	90.55	90.84	90.52	90.53
Our	Cross-modal Fusion (Proposed)	**91.61**	**91.92**	**91.46**	**91.53**
	DistilBERT (Baseline)	**91.59**	**91.71**	**91.44**	**91.49**

D. Computational Cost

The fusion model's 433.2M parameters make it roughly 6.5× larger than the 66.9M-parameter DistilBERT baseline, translating into proportionally higher memory footprint and per-batch latency, primarily driven by the additional CLIP vision tower and cross-attention layers. Given that the accuracy gain of the fusion model over DistilBERT on this text-only benchmark is small (≤0.3 pp across metrics), DistilBERT represents a more efficient choice when only text is available; the fusion architecture's value proposition instead lies in its native support for image-bearing posts, where its additional capacity is expected to be more decisively exploited. This cost/accuracy trade-off should be weighed against deployment constraints (e.g., real-time moderation at scale) when selecting between the two architectures in production.

E. Implication and limitation

The proposed framework gives a flexible solution for cyberbullying detection by combining transformer-based text representation with cross-modal attention and gated fusion. However, the main limitation is the use of a text-only dataset, which restricts proper validation of the visual component. The model also has relatively high computational requirements due to its large architecture. Future work should evaluate the framework on genuine text-image datasets, improve computational efficiency, and explore

stronger multimodal and explainable learning techniques.

V. Conclusion and Future work

The proposed Cross-modal Fusion model demonstrates effective cyberbullying detection by integrating XLM-RoBERTa text features, CLIP visual representations, BCAM, and gated feature fusion. The model achieves 91.61% accuracy and 91.53% macro-F1, slightly outperforming the DistilBERT baseline. The explainability analysis using SHAP and LIME further provides insight into the textual features influencing model decisions. However, the current evaluation is limited by the text-only nature of the dataset, restricting validation of the visual modality. In future work, the framework will be evaluated on additional publicly available multimodal cyberbullying datasets, such as MMHS150K, to validate its performance on genuine text-image content. Further experiments will investigate advanced multimodal architectures, including ViLT, VisualBERT, and CLIP-based transformer fusion models, along with improved attention and fusion strategies. Future studies will also focus on reducing computational complexity and enhancing explainability to develop a more robust and practical cyberbullying identification system.

References

[1] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," Mathematics, 2023, doi: 10.3390/math11163567.

[2] F. R. Sayed, E. H. Elnashar, and F. A. Omara, "Cyberbullying detection in social media using natural language processing," 2025. doi: 10.1016/j.sciaf.2025.e02713.

[3] A. G. Philipo, D. Sebastian Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Assessing Text Classification Methods for Cyberbullying Detection on Social Media Platforms," IEEE Trans. Inf. Forensics Secur., 2025, doi: 10.1109/TIFS.2025.3588728.

[4] C. Van Hee et al., "Automatic detection of cyberbullying in social media text," PLoS One, 2018, doi: 10.1371/journal.pone.0203794.

[5] B. M. Yakubu, W. Bumrungsri, and P. Bhattarakosol, "Natural Language Processing (NLP) System for Cyberbullying Identification in Thai Language," J. Comput. Cogn. Eng., 2025, doi: 10.47852/bonviewJCCE52024355.

[6] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Adv. Sci. Technol. Eng. Syst., 2017, doi: 10.25046/aj020634.

[7] B. Ogunleye and B. Dharmaraj, "The Use of a Large Language Model for Cyberbullying Detection," Analytics, 2023, doi: 10.3390/analytics2030038.

[8] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," Online Soc. Networks Media, 2023, doi: 10.1016/j.osnem.2023.100250.

[9] Y. Kumar et al., "Bias and Cyberbullying Detection and Data Generation Using Transformer Artificial Intelligence Models and Top Large Language Models," Electron., 2024, doi: 10.3390/electronics13173431.

[10] S. Chen, J. Wang, and K. He, "Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model," Inf., 2024, doi: 10.3390/info15020093.

[11] J. Cao, Y. Xiong, W. Wang, and G. Chen, "Cyberbullying Detection Based on Hybrid Neural Networks and Multi-Feature Fusion," Information, vol. 17, no. 2, p. 205, Feb. 2026, doi: 10.3390/info17020205.

[12] T. D. Pal, D. D. Kondekar, and F. khan T. Khan, "Cyberbullying Detection on Social Media Using Machine Learning," Int. J. Innov. Res. Technol., vol. 12, no. 9, pp. 1–9, 2026, [Online]. Available: https://ijirt.org/publishedpaper/IJIRT192334_PAPER.pdf

[13] C. Kumari and M. Kaur, "Towards Enhanced Cyberbullying Detection: A Unified Framework with Transfer and Federated Learning," Systems, vol. 13, no. 9, p. 818, Sep. 2025, doi: 10.3390/systems13090818.

[14] S. Garc and F. De Arriba-p, "Promoting security and trust on social networks : Explainable cyberbullying detection using Large Language Models in a stream-based Machine Learning framework," arXiv

Comput. Sci., no. October, 2024.

[15] A. Muneer, A. Alwadain, M. G. Ragab, and A. Alqushaibi, "Cyberbullying Detection on Social Media Using Stacking Ensemble Learning and Enhanced BERT," Information, vol. 14, no. 8, p. 467, Aug. 2023, doi: 10.3390/info14080467.

[16] A. Dewani et al., "Detection of Cyberbullying Patterns in Low Resource Colloquial Roman Urdu Microtext using Natural Language Processing, Machine Learning, and Ensemble Techniques," Appl. Sci., vol. 13, no. 4, p. 2062, Feb. 2023, doi: 10.3390/app13042062.

[17] H. Basheer, "Cyber Bullying Detection in Twitter using Machine Learning Algorithms," Comput. Sci., 2022, doi: https://doi.org/10.35629/5252-040811721184.

[18] M. Alotaibi, B. Alotaibi, and A. Razaque, "A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media," Electronics, vol. 10, no. 21, p. 2664, Oct. 2021, doi: 10.3390/electronics10212664.