International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 1200-1217
https://doi.org/10.70917/ijcisim-2026-5388
Article

Uncertainty Calibrated Cross Platform Identity and Dyadic Episode Modeling for Cyberbullying Analysis

Sathea Sree.S[1], L. Nalini Joseph[2]

[1]Research Scholar, Department of Computer Science and Engg, Bharath Institute of Higher Education and Research, Chennai. Email: Satheasadasivam@gmail.com
[2]Professor, Department of Computer Science and Engg, Bharath Institute of Higher Education and Research, Chennai. Email: nalinijoseph23@gmail.com

Abstract: Cyberbullying is an important issue in digital safety because hostile communications remain accessible on the Internet, intensify over time, and can have profound psychological and social impacts. Timely and proportional moderation is therefore crucial and cannot be achieved without accurate automated detection. Existing studies rely on attention networks, recurrent networks, contextual embeddings, temporal fusion, graph networks, and metaheuristic optimization, but the majority take only single-stage text content into consideration and fail to collectively verify the repeated targeting behavior, cross-platform identity consistency, dynamic power imbalance, harm-directedness, and uncertainty of prediction. To fill this gap, this study presents Causal Identity-aware Dyadic Episode Reasoning for Cyberbullying (CIDER-CB), an advanced learning framework that contains three phases. Phase I conducts multimodal representation with reliability-gating, temporal heterogeneous-graph construction, and probabilistic cross-platform identity alignment. Phase II deploys a marked temporal point process to measure repetition and escalation, a graph-based dyadic reasoning approach to measure the power asymmetry between perpetrator and victim dyad components, and counterfactual causal learning to measure harm towards a target. Phase III combines fairness-constrained intervention learning, evidential uncertainty estimation, and concept drift-aware continual adaptation in order to allow for reliable real-time moderation systems. Through the episode-level and empirical trials, the approach attains effective outcomes via user-disjoint, temporally-separated validation. Thus, the framework achieves 96.5% recall, 98.5% total accuracy, 97.0% precision, 96.7% F1-score, and less than 2.0% false-positive rate. All of these attainments exhibit the proposed performance targets.
Keywords: cyberbullying, episode, dyadic, uncertainty, accuracy, cross-platform, harm, behavior, victim, detection

1. Introduction

Cyberbullying is a form of bullying that is expressed through the repeated use of electronic communication devices/approaches and is deliberate and causes significant harm to the victim who is unable to retaliate to the same extent as the aggressor [1]. It has been defined as cyber-aggression that involves three elements: intentional, persistence, or directed harm, and power imbalance, which distinguishes it from isolated cyber-aggression [2]. These properties are exacerbated in the online world, as the material may be archived and continue to be available, may be shared to a larger audience, may be reposted as replies and mentions, and may be shared in various accounts or on different platforms. An accurate detection system therefore needs to be able to assess, as well as look for offensive language, who is targeting whom, repeatability, the chronology of the interaction, the social and/or structural disadvantage of the target, and the adequacy of the available evidence for moderation [3].

Current research on cyberbullying has shifted from the traditional lexical and machine-learning classifiers to contextual transformers, conversation-level modeling, temporal graph neural networks, bystander role analysis, continuous-time forecasting, multimodal fusion, and AI-based intervention for cyberbullying [4]. Fine-grained models have shown beneficial incorporation of conversational dynamics and bystander dynamics, with temporal social-graph

approaches adding interest in modeling temporality and sparse users, and studies on generative Artificial Intelligence (AI) further pointing to growing interest in context-sensitive intervention versus detection [5]. Still, such gains are going undervalued. Even the majority of the approaches fail to account for identity continuity across platforms and repeated targeting from perpetrators to victims, as well as escalation, dynamic power inequality, prediction of harm or group fairness, target-specific harm, and proportionate intervention [6]. Hence, an incident with a high score for toxicity can still be classified as cyberbullying, even in the absence of repetition or power imbalance, and harmful incidents can happen that are spread across formats and platforms without detection [7].

To overcome this limitation, this study introduces Causal Identity-Aware Dyadic Episode Reasoning for Cyberbullying (CIDER-CB), a novel approach, which reformulate cyberbullying detection as the verification of a temporally evolving episode of social structure, instead of verifying a single message. CIDER includes three phases. Phase-I includes the identity-aware multimodal event grounding strategy that fuses and represents multimodal information (heterogeneous text, metadata, image, audio, profile, meme, video, and interaction signals) into event representations and assigns reliability information to each. Thus, probabilistically aligns platform-specific accounts, and builds an identity-consistent temporal heterogeneous graph. Phase II comprises causal dyadic episode reasoning, which identifies and measures repetition and escalation with the marked temporal point process, dynamic power imbalance with the graph-based method, and the specific target of the harmful content with the counterfactual analysis. Phase III involves uncertainty-calibrated adaptive intervention, aimed at estimating the degree of evidential uncertainty, sending ambiguous cases back to humans for review, and selecting ratioed actions for fairness, utility, and concept drift constraints.

1.1 Contribution of the Study

The study has three inter-related technical contributions. Reliability-gated multimodal event grounding and probabilistic identity association are introduced as first-order, which are capable of representing the interaction history as a coherent behavioral history without any assumption on deterministic identity matches, from the perspective of multiple events, since they can come from multiple accounts and platforms. Second, it builds the episode-level causal reasoning mechanism where repetition is treated as the complement definitional condition of target-directed harm and power imbalance; this weighted conjunction is a strategy to avoid a strong signal of toxicity offsetting the lack of a complementary signal. Third, it maps the detection to evidential uncertainty [8], fair limits to the policy-learning process, and drift-detection to adaptation, to turn the model output from a rigid label to a moderation decision that is sensitive to evidential uncertainty and constrained by concepts of fairness, while at the same time being sensitive to the detection of drift.

1.2 Research Novelty

The key novelty of CIDER-CB is simultaneously combining identity uncertainty, multimodal evidence, counterfactual harm-directedness, continuous-time dyadic escalation, and dynamic power imbalance and fairness-aware intervention in a single episode-verification framework. While many systems might have multimodal representations, transformers, or temporal graphs, they do not necessarily have participant roles, nor intervention generation; CIDER-CB is designed to analyze whether sets of cross-platform events meet the defining structure of cyberbullying. So, it is not any other feature extractor and/or classifiers, but a behavioral and identity-aware reasoning architecture that connects evidence grounding, uncertainty calibration, definitional verification, and intervention selection.

1.3 Scope and Motivation

The scope of the study covers multimodal and multi-platform social-media events organized into temporally ordered perpetrator–victim episodes. It addresses binary-episode identification, probability calibration, early detection, severity estimation, and audit-group fairness and intervention selection under varying online language and interaction patterns. The study is driven by the necessity of operations to objectify early identification of “true” bullying with a minimization of false accusations and disproportionate moderation and omissions of incidents of

severe bullying. It is not a diagnostic tool to understand the psychological state of users, to determine real-world identities behind anonymous accounts, or to automate human moderation decision-making in ethically grey situations, but rather to give structured evidence, calibrated risk estimates, and intervention suggestions, which will help to create safer and more accountable moderation.

1.4 Objectives

The following technical description exhibits the major objectives of the study:

- To create an identity-aware multimodal event-grounding approach which reliability weights heterogeneous data and probabilistically connects the interactions on various platforms within two identity-centered temporal interaction graphs.
- To develop a causal dyadic episode-reasoning model of repetition and escalation, target-specific harm-directedness and dynamic perpetrator–victim power imbalance.

The rest of the manuscript is structured as follows. In Section 2, related cyberbullying studies are presented, spanning from post-level classifiers, session-level classifiers, multimodal learning, conversational and bystander modelling, temporal graph modelling, causal reasoning, uncertainty estimation, to systems addressing intervention-oriented. The research gap is identified at the end of this section. Section 3 presents the complete CIDER-CB methodology through identity-aware multimodal event grounding, causal dyadic episode reasoning and uncertainty-calibrated adaptive intervention. In section 4, the dataset, the implementation environment, the comparative approaches, the hyperparameters and the evaluation measures are described. Section 5 presents the calibration, fairness, early-detection, and significance-tested results technically. And also addresses implications, limitations and threats to validity. Section 6 wraps up the study and presents directions for future work in the areas of cross-platform deployment, expansion of demographic auditing, and human-in-the-loop moderation.

2. RELATED STUDIES

Several existing approaches and strategies are discussed in this section for precise understanding of the proposed study. Authors in [10] presented BullySpace, an analogySpace common-sense reasoning framework, which combines social-network information, natural-language processing, supervised text classification, domain-specific common-sense knowledge, and reflective intervention interfaces. BullySpace encoded over 200 statements about bullying that are stereotypical, but these were mixed with ConceptNet and fed into AnalogySpace's singular-value-decomposition reasoning mechanism to identify subtle attacks that may not contain any profanity. The three annotators agreed with the model's reasoning on 70%, 76%, and 68% of the 50 cases of implicit Formspring bullying. High in-context positive agreement for bystander reflection (100%), victim helpfulness (100%), and perpetrator self-reflection (80%) further demonstrated its effectiveness compared to static assistance.

Researchers in [11] presented a Multi-Task and Multi-Modal (MT-MM) framework for sentiment- and emotion-aided cyberbullying detection in the Hindi–English code-mixed tweets. The researchers created the 6,084-instance BullySentEmo corpus, which includes tweet text, emojis, cyberbullying labels, three sentiment classes, and seven emotion classes. Bidirectional GRU and attention layers learn context-sensitive features, and cyberbullying detection is considered the main task, while sentiment analysis and emotion recognition are considered secondary tasks, where the shared multi-task learning approach is used. The model achieved an accuracy of 82.05% ± 1.36% for cyberbullying detection, 77.87% ± 1.93% for sentiment analysis, and 58.05% ± 2.78% for emotion recognition, which is higher than the single-task and unimodal models evaluated.

The study by [12] employed a role-based qualitative content-analysis methodology, in which the Twitter discourse of three roles related to bullying (target, perpetrator, helper) was analyzed. The procedure retrieved and cleaned 847,548 tweets, introduced TF–IDF representation and classified tweets as bullying traces using logistic-regression and support-vector-machine classifiers, and then conducted detailed qualitative coding of 780 tweets (260 tweets for each of the roles). A first screening process found that 240,018 of the 240,018 tweets being processed had

evidence of bullying, corresponding to 28.58%, and the procedure of qualitative coding achieved greater than 80% intercoder agreement. Results indicated that 67.0% of the target tweets mentioned past bullying, 65.76% of the perpetrators knew their targets, and helper tweets mainly were reporting (31.52%), defending (36.23%), and accusing (26.81%).

To explore disclosures of cyberbullying on Twitter, the study in [13] used a Hybrid Supervised Machine Learning, and Qualitative Content Analysis (HSML-QCA) approach, without giving a specific acronym to it. The tweets were gathered using primary, and secondary keywords related to bullying; spam-like hashtags, political references, and non-English tweets were removed; tokenization, part-of-speech tagging, lemmatization, logistic regression, unigram–bigram representations, and Support Vector Machine (SVM) were applied. The classifiers first identified the traces of bullying and then isolated 69,963 cyberbullying-related tweets, followed by an examination of author roles and purposes, and a qualitative content analysis of 500 tweets examined victim-bully relationships, the places of the incidents, the type of behavior and the time frame. The bullying-trace classifier achieved 70% accuracy, and the following cyberbullying classifier achieved 78% accuracy, with a naïve baseline accuracy rate of 34%, with 28.58% of the larger tweet set identified as bullying traces.

Investigators in [14], proposed a two-stage deep-learning architecture for the Chained Fine-Grained Cyberbullying Detection (CFGC) model with Bystander Dynamics that adds full context from the Twitter-thread. The first classification layer uses the fine-tuned BERT model to predict bystander roles (instigator, impartial, defender, or other), and the roles are used as the dependent features for the second layer to classify four different types of communicative situations (normal communication, aggression with no bullying, bullying with low aggression, and bullying with high aggression) using a classifier chain with a chain of two LSTMs. The method thus represents the linkage between bystander reactions and the severity of cyberbullying, rather than as a binary post for each case. The incorporation of the bystander roles increased performance by 25.35%, and the final model gave an 89% weighted F1-score with a 95% confidence interval (78%, 97%).

Authors in [15] developed a Multi-Agent Reinforced Guided Weighted Multi-Relational Subgraph Neural Network (RSBully) for multimodal cyberbullying detection. The framework initially models the heterogeneous information from social media components, including posts, entities, users, images, locations, and session information, as a Heterogeneous Information Network (HINet), and then transforms it into a weighted multi-relational graph. A reinforcement-learning mechanism is used to determine aggregation thresholds for each relation is used to screen the subgraphs for structural importance, eliminate redundant information and capture intrinsic relationships between modalities and bullying behavior. The chosen examples are combined to represent social media sessions and offer some structural interpretability of influential entities and relationships. Experiments conducted on real-world social-media datasets demonstrated that RSBully always performed better than the evaluated state-of-the-art models.

In order to address the challenges of low network connectivity and few labels, the researchers in [16] developed a Temporal Social-Graph learning model, called TemSoGraph, which was designed to detect and predict cyberbullying. The framework is designed to map the user interactions into a graph structure that dynamically evolves as users interact with the content, incorporate temporal self-attention to model both short-term and long-term patterns of user interactions, simulate low-activity users by jointly updating the local user representation and global graph representation, and utilize a domain adaptor to transfer features across platforms. It was measured with Instagram data that includes 7,180 users and 37,808 posts, and Vine data that consists of 6,816 users and 78,599 posts. When it comes to future comment-level prediction of cyberbullying, TemSoGraph achieved excellent recall rates (97.18% on Instagram and 93.38% on Vine), F1-scores (97.11% on Instagram and 92.86% on Vine) with an AUC score (99.57% on Instagram and 97.05% on Vine).

The research work in [17] presents a psychologically based model for generating bystander responses to cyber aggression, which was developed using a Large Language Model (LLM) called AllyGPT. This method first analyzes the level of incivility of the aggression and whether the person or group is being attacked, and then decides on one of five intervention strategies: calling out aggression, advising disengagement, expressing empathy, correcting

misinformation, or redirecting discussion. The researchers created GPT-4o responses for 996 cyberaggression scenarios, which were evaluated using the linguistic analysis, human ratings, and qualitative interviews, with three types of responses created: policy-reminder, baseline-GPT, and AllyGPT. AllyGPT gained a mean rating of 3.89/5 for discussion alteration, 3.98/5 for favorability, and 3.45/5 for behavioral change, with normalized attainments of 77.8%, 79.6%, and 69.0%, respectively. While AllyGPT showed relatively positive performance on low-incivility cases, the performance on the other cases was below the baseline GPT.

In contrast to a predictive cyberbullying classifier, authors in [18] created an integrated Bystander Intervention Model-Social Ecological Structural Equation framework to account for the relationship between cyberbullying-perpetration experience and bystander behaviors. A total of 549 valid responses of university students were analyzed and five types of bystander responses, social relationship closeness, and moral disengagement were measured. A confirmatory factor analysis was followed by structural equation modeling with 5,000 iterations of bootstrap mediation and moderation testing. Moral disengagement mediated the influence of perpetration experience on negative bystander behavior to explain 49% of the influence on reinforcing the bully and 44% of the influence on assisting the bully. The mediation model also fit well, with a Comparative Fit Index of 92.4% and Tucker-Lewis Index of 91.7%; higher social relationship closeness had a negative effect on perpetrators' tendencies to assist and reinforce perpetrators.

Current methods predominantly focus on post/session categorization, and tend to neglect temporal escalation, repeated dyadic harm, power imbalance, and cross-platform identity continuity. Adaptive intervention and limited uncertainty calibration, fairness control, and early-detection assessment are also provided regardless of multimodal use, graph features, or bystander use. Hence, CIDER-CB models the following events in a multimodal way, while considering them in the context of identity: (i) evolution of causal events, (ii) fairness-constrained decisions, (iii) calibrated uncertainty, and (iv) timely intervention.

3. METHODOLOGY

The overall architecture of the Causal Identity-Aware Dyadic Episode Reasoning for Cyberbullying (CIDER-CB) framework is depicted in Figure 1, which aims to take cyberbullying analysis beyond the classification of individual toxic posts into the verification of dyadic and socially structured bullying episodes. The architecture commences with the heterogeneous multi-platform input feeds, which are text, audio, memes/images, timestamps, video, user profiles, metadata, and interactions. In Phase I, modality-specific encoders encode these inputs into modality specific semantic representations, and a reliability-gated multimodal fusion suppresses noisy or missing signals, with relations linked with confidence-aware associations, while a dynamic heterogeneous interaction graph organizes posts, users, platforms, and relational events into identity-consistent temporal structure. Phase II implements causal dyadic episode reasoning, escalation (marked temporal point process), counterfactual analysis (was it really about the target, mark inflection), consisting of repetition, and dynamic estimation of power imbalance (did it amount to cyberbullying, mark inflection), with a definitional conjunction producing episode probability and episode severity. Phase III is based on an evidential uncertainty estimation and selection of an appropriate intervention (fairness constraints for victims), drift-aware continual learning, allowing for monitoring, human review, warning prompt, visibility reduction, victim support, safety escalation, or temporary restriction. The feedback pathway also enables the system to revise previous representations and reasoning modules according to the changes in the patterns and the behaviors of language and interaction. The following methodology section briefs the core phases of the CIDER-CB working mechanism.

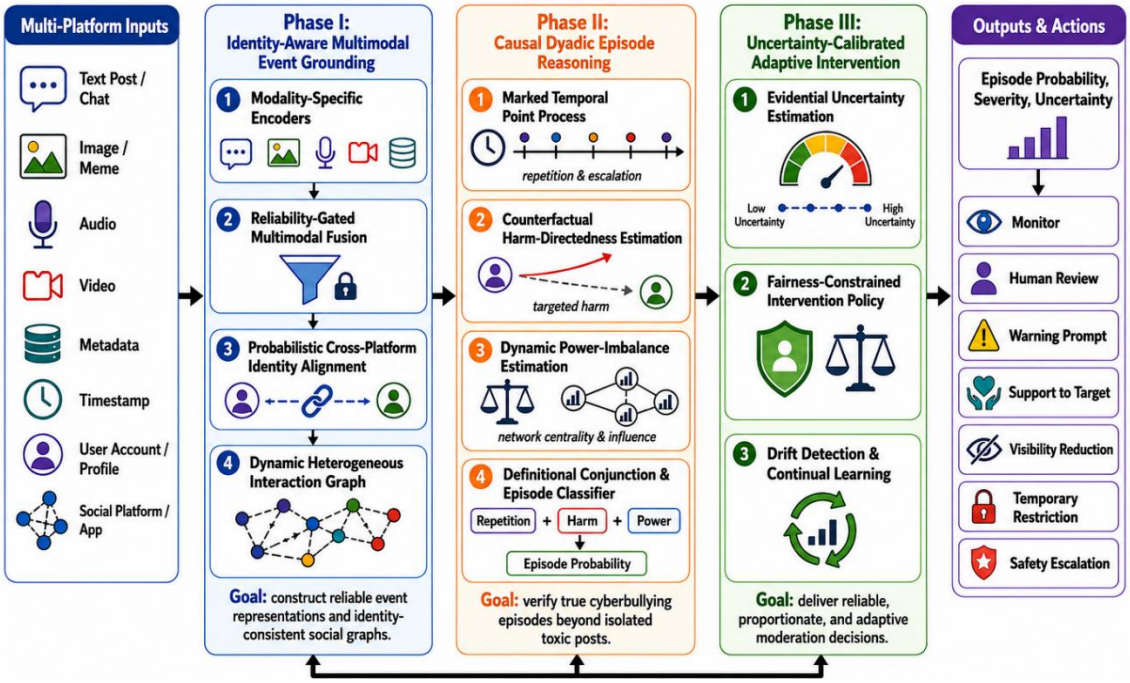

Figure 1. Architectural overview of Causal Identity-Aware Dyadic Episode Reasoning for Cyberbullying Framework

3.1 Identity-Aware Multimodal Event Grounding

The first phase maps observations of social media events (which are noisy and diverse in nature) to coherent and factually consistent interaction objects. It is critical because cyberbullying might be delivered via memes, text, audio, pictures, video, reply designs, emojis, and involvement actions, and a particular actor could maintain multiple accounts on a similar system or use completely different accounts on totally different systems. Simultaneously examining each separate instance of the modality or harassment account can lead to an untargeted continuing episode, and a failure of the model to capture an instance of repeat targeting. To achieve this, at first modality-specific encoders extract the semantic information and then the reliability gates assign higher confidence to the informative modalities and lower confidence to modalities such as the ones that were missing, corrupted, or did not have high confidence. Such unified event representation (U_i) is computationally expressed as,

$$U_i = \sum_{m \in M_i} \quad \varepsilon_m\left(I_i^{(m)}\right) \cdot F_m \cdot L_i^{(m)} + \exists_{P_i} + T(t_i),\ L_i^{(m)} = \frac{exp\left(r_i^{(m)}\right)}{\Sigma_{(a \in M_i)} \quad exp\left(r_i^{(l)}\right)} \qquad \textbf{(1)}$$

In eq. (1), i represent denotes a social-media event, M_i denote the set of available modalities; $I_i^{(m)}$ specifies the input from modality m, $\varepsilon_m(\cdot)$ signifies its modality-specific encoder, F_m points the modality features (into a shared space), $L_i^{(m)}$ indicates learned reliability weight, $r_i^{(m)}$ denotes the reliability score derived from encoder confidence and feature quality, $\exists_{P_i}$ signifies the platform embedding, $T(t_i)$ indicate the temporal encoding and a denotes the modality I available in M_i.

The second core process is to do probabilistic identity alignment across platforms and dynamically construct graphs. The model avoids making any assumptions about the identity of two accounts being the same, but instead estimates the probability of the accounts belonging to the same identity based on linguistic style, interaction neighborhoods, temporal activity patterns, and permissible profile evidence. These probabilities are added to a temporal heterogeneous graph of users, posts, platforms, topics, replies, mentions, attacks, support given and received.

This process is crucial in resolving the history of perpetrator-victim relationships and is also for reducing erroneous combinations of accounts. Thus, the optimal identity-association matrix is constructed as,

$$|O|^{*} = arg\ arg\ \min_{O \in \Upsilon(k)} \left[\sum_{(\alpha,\beta)} \left(\mathsf{C}_{(\alpha\beta)} \cdot O_{ab} \right) + \epsilon \sum_{(\alpha,\beta)} O_{ab} (log\ log\ O_{ab} - 1) \right] \quad \textbf{(2)}$$

The accounts in this case of eq. (2) are denoted with α and β. Further, O_{ab} represents the estimated probability mass between two accounts, O denote the optimal identity-association matrix, $\mathsf{C}_{(\alpha\beta)}$ is the account-matching cost calculated based on differences in behavioral, temporal, linguistic, network, and authorized profiles, $\Upsilon(k)$ represent the set of transport matrices satisfying account-distribution marginals j and k, and ϵ signifies the entropy regularization, which ensures that identity matching is stable and non-deterministic.

3.2 Casual Dyadic Episode Reasoning

The second phase refers to whether a sequence was a process that can be considered a real cyberbullying episode or an isolated offensive post. The first process is a marked temporal point process that quantifies the frequency of acts of aggression directed towards a target by a candidate perpetrator, whether or not they occur in meaningful time intervals, and whether their severity is increasing. This is essential as one of the fundamental characteristics of cyberbullying is that they are repeated acts and conventional classifiers take each of these acts as an independent observation. Event marks can be signs of insult, humiliation, threat, coercion, exclusion, neutral communication, or sexual harassment. Thus, the conditional intensity of a future event, $\delta_{\rho v}^{(k)}$ of type k (interaction mark) is computationally defined as,

$$\delta_{\rho v}^{(k)}(t) = \left[x_k + G_{\rho v}(t) \cdot q_k^{\top} + \sum_{\mathrm{e}_j \in h_{\rho v}(t)} \lambda_{k,\Theta_j} \cdot exp\ exp\ \left(-\Upsilon_k \left(t - t_j \right) \right) \right] \quad \textbf{(3)}$$

In eq. (3), ρ and v denotes the candidate and victim perpetrator, respectively. t denotes the current time, x_k, and q_k signifies the learned parameters, $G_{\rho v}$ represent the present dyadic graph representation, $h_{\rho v}$ reveals the historic interaction preceding at any given t, Θ_j and t_j indicate the mark and time of the historical event e_j, respectively, λ_{k,Θ_j} denote the measure of to what extent a historical Θ_j excites a future k, Υ_k implies the temporal decay.

Further, the probability of repeated or persistent aggression, $r_{\rho v}$ is computed as,

$$r_{\rho v}(t) = 1 - exp\ exp\ \left[- \int_{t-\varpi}^{t} \sum_{k \in \ddot{a}} \lambda_{\rho v}^{(k)}(\eta)\, d\eta \right] \quad \textbf{(4)}$$

In eq. (4), $\ddot{a}$ indicate the set of aggressive marks, ϖ signifies the observation window, and η projects the continuous time variable within $[t - \varpi, t]$.

The second process of this phase collaborated on the estimation of harm-directedness and dynamic power imbalance. Counterfactual reasoning helps to distinguish between targeted (aggression) and general (dirty word, topic discussion, sarcasm, quotation) usage of aggressive cue words by considering whether removing words that target the specific individual would significantly alter the degree of harm predicted (as opposed to the general form of aggressive cue words). Power imbalance is deduced from the network centrality, audience size, reply dominance, past sanctions, coalition, and reciprocity, and interactional vulnerability of the target. It is important to take these variables into consideration since repetitive, antagonistic interactions between two or more users with equal participation might not be an instance of cyberbullying but a manifestation of conflict. As a result, the final episode decision is based on a definitional conjunction that requires that repetition, directed harm, and power imbalance gain sufficient support. Therefore, the predicted probability of a cyberbullying episode at any given t, $\hat{P}_{\rho v}^{CB}(t)$ is computed as,

$$\hat{P}_{\rho v}^{CB}(t) = \left[\left(\varphi_{\rho v}(t)^{\omega_\varphi} \right) \cdot \left(r_{\rho v}(t)^{\omega_r} \right) \cdot \left(\pi_{\rho v}(t)^{\omega_\pi} \right) \right]^{\left(\omega_\varphi + \omega_r + \omega_\pi \right)^{-1}} \cdot \Sigma \left(\theta_{\rho v}^{epi}(t) \cdot x_c^{\top} + y_c \right) \quad \textbf{(5)}$$

In eq. (5), $r_{\rho v}(t)$, $\varphi_{\rho v}(t)$, and $\pi_{\rho v}(t)$ projects the repetition score, the counterfactual harm-directedness score, and the dynamic power-imbalance score, respectively, $\theta_{\rho v}^{epi}(t)$ signifies the overall episode exhibitor containing graph,

temporal semantic, and escalation aspects. Further, ω_r, ω_φ, and ω_π denotes positive significant parameters, x_c and y_c represent the classification parameters; and $\Sigma(\cdot)$ signifies the sigmoid operator. The weighted geometric conjunction means that the presence of a strong signal cannot completely offset the lack of another key signal.

3.3 Uncertainty-Calibrated Adaptive Intervention

The third phase is the step to translate episode predictions to proportional and reliable moderation decisions. Its first process is the evidential uncertainty estimation, where the network predicts both class probabilities and the evidence amounts supporting them. This is key as ambiguous humor, cultural expression, use of reclaimed language, the ambiguous identity relationship, and incomplete history can give rise to seemingly sure, albeit wrong inferences in classical softmax classifiers. The evidential representation enables the machine not to execute an automatic modification and passes on indeterminate cases to a human moderator.

$$e_s = softplus\left(f_s\left(\theta_{pv}^{epi}\right)\right);\ \Omega_s = e_s + 1;\ \hat{P}_s = \left(\Omega_s \cdot \left(\sum_{j=1}^{K} \quad \Omega_j\right)^{-1}\right),\ \psi = \left(K \cdot \left(\sum_{j=1}^{K} \quad \Omega_j\right)^{-1}\right) \quad \textbf{(6)}$$

In eq. (6), s denotes an episode-severity class, K denote the number of classes; $f_s(\cdot)$ denotes the class-specific evidence function, e_s denotes non-negative evidence for severity class s; Ω_s projects the respective Dirichlet concentration parameters; $\hat{P}_s$ denotes the expected probability of class severity, and ψ indicate the overall predictive uncertainty. The accumulated evidence, if it is large, leads to low uncertainty; if it is minimal or conflicting, implies high uncertainty and will be passed on to human analysis.

During adaptation to online language change, the second core process uses fairness and utility restrictions to choose interventions through policy learning. Before selecting the actions of continued monitoring, visibility reduction, reconsideration prompt, temporary restrictions, victim-support resources, and human safety escalation, the policy takes into account episode probability, uncertainty, severity, prior interventions, escalation, and contextual risk. This process is important as the detector does not reveal which of the responses will minimize harm and a highly automated policy might have a disproportionate impact on specific language or demographic groups. Thus, the optimization seeks to reward avoidance of harm and to punish false alarms, the cost of operation, the number of undetected severe avoiding events, and the imbalance across the group.

$$\nabla^* = arg\ arg\ \max_{\nabla}\ E_\nabla\left[\sum_{t=0}^{T} \quad \left(\mathrm{R}_t^H - \Gamma(a_t) - \mu_{FN} \cdot D_{(FN,t)} - \mu_{FP} \cdot D_{(FP,t)}\right)\gamma^t\right] \quad \textbf{(7)}$$

In eq. (7), γ implies a temporal discount factor, R_t^H indicate the estimated value of the reward for the prevented harm, a_t signifies the selected intervention, $\Gamma(a_t)$ defines the operational and social cost of a_t, $D_{(FN,t)}$ and $D_{(FP,t)}$ indicate a false-negative and false-positive decision, respectively, μ_{FN} and μ_{FP} estimate their penalties, and finally, t and T indicate the decision times and the decision horizon, respectively. The execution is subjected to the comparison group $\left(z^{'}\right)$ and maximum allowed unfairness gap (N) where false positive rate $\left(FPR_z\right)$ and true positive rate $\left(TPR_z\right)$ are evaluated for z.

In a sequence, the following CIDER-CB algorithm, consisting of multimodal fusion with reliability-gated, temporal dyadic reasoning, probabilistic identity alignment, and cyberbullying episode estimation, is performed. It then calculates severity and uncertainty, and chooses a moderation intervention that is within the bounds of fairness.

Algorithm: Working Procedure of CIDER-CB

Input: $D = \{t_i\}_{i=1}^{N},\ m \in M_i$ **Output:** $\{\psi a_t\}$
Step 1: Initialize $i \leftarrow 1,\ t \leftarrow 0$ **Step 2: FOR** $i = 1$ **to** N **Encode–Gate–Fuse**

$$L_i^{(m)} = \frac{exp\ exp\ \left(r_i^{(m)}\right)}{\left(r_i^{(l)}\right)} U_i = \sum_{m \in M_i} \varepsilon_m \left(I_i^{(m)}\right) \cdot F_m \cdot L_i^{(m)} + \exists_{P_i} + T(t_i)$$

Store

$$U_i \longrightarrow multimodal\ event\ set$$

END FOR

Step 3: FOR each cross-platform account pair (β)

Compute $\mathsf{C}_{\alpha\beta}$

Optimize Identity Association

$$|O|^* = arg\ arg \min_{O \in \mathfrak{l}(k)} \left[\sum_{(\alpha,\beta)} \left(\mathsf{C}_{(\alpha\beta)} \cdot O_{ab}\right) + \epsilon \sum_{(\alpha,\beta)} O_{ab}(log\ log\ O_{ab} - 1) \right]$$

Update temporal heterogeneous graph using U_i **and** O^*

END FOR

Step 4: FOR each active dyad (v)

Extract $G_{\rho v}(t),\ h_{\rho v}(t)$

FOR each interaction mark k

$$\delta_{\rho v}^{(k)}(t) = \left[x_k + G_{\rho v}(t) \cdot q_k^{\top} + \sum_{\mathrm{e}_j \in h_{\rho v}(t)} \lambda_{k,\theta_j} \cdot exp\ exp\ \left(-\Upsilon_k\left(t - t_j\right)\right) \right]$$

END FOR

Compute Repetition

$$r_{\rho v}(t) = 1 - exp\ exp\left[- \int_{t-\varpi}^{t} \sum_{k \in \ddot{a}} \lambda_{\rho v}^{(k)}(\eta)\, d\eta \right]$$

Estimate $\varphi_{\rho v}(t), \pi_{\rho v}(t)$

Construct $\theta_{\rho v}^{epi}(t)$ //episode representation
Compute $\hat{P}_{\rho v}^{CB}(t)$ //Cyberbullying Episode Probability

Step 5: FOR $s = 1$**to** K

Compute Evidence

$$e_s = \left[f_s\left(\theta_{\rho v}^{epi}\right)\right] \Omega_s = e_s + 1$$

Compute Severity Probability, $\hat{P}_s = \Omega_s\left(\sum_{j=1}^{K} \Omega_j\right)^{-1}$
END FOR

Compute Uncertainty, $\psi = \left(K \cdot \left(\sum_{j=1}^{K} \Omega_j\right)^{-1}\right)$

Step 6: Optimize Intervention Policy

$$\nabla^* = arg\ arg\ \max_{\nabla} E_{\nabla}\left[\sum_{t=0}^{T} \left(\mathrm{R}_t^H - \Gamma(a_t) - \mu_{FN} \cdot D_{(FN,t)} - \mu_{FP} \cdot D_{(FP,t)}\right)\gamma^t\right]$$

subject to:

$$| FPR_z - FPR_{z'} | \leq N \mid TPR_z - TPR_{z'} | \leq N$$

Step 7: Select $a_t \leftarrow \nabla^*(\hat{P}_s \psi)$
Step 8: Update–Repeat

$$t \leftarrow t + 1$$

$IF\ t \leq T\ GOTO$ Step 2

$$ELSE\ RETURN\ \{\psi a_t\}$$

4. MATERIALS EMPLOYED

The empirical implementation of CIDER-CB was performed with Ubuntu 22.04.4 LTS, Python v3.11.9, PyTorch Geometric v2.5.3, pandas v2.2.2, PyTorch v2.3.1, scikit-learn v1.5.1, Hugging Face Transformers v4.42.4, and NumPy v1.26.4, with support for GPU acceleration in the form of CUDA v12.1 and cuDNN v9.1. The experiments were conducted on a workstation with an i9 processor (Intel Core) and 64 GB of memory (DDR5 RAM), which were able to provide sufficient computational power for multimodal encoding, episode-level training, temporal graph processing, repeated evaluation, and uncertainty estimation.

4.1 Dataset

For this study, the Temporal Multimodal Cyberbullying dataset (TMC dataset) from [21] was taken into consideration for evaluating the study due to the fact that the dataset directly satisfies the CIDER-CB requirements of multimodal event grounding, cross-platform identity consistency, dyadic temporal reasoning, repeated-target detection, harm directedness analysis, dynamic power-imbalance estimation, uncertainty calibration, and intervention assessment. Each social-media event in the dataset is chronologically ordered and exists as a dyadic episode, comprising around 10,248 events with all essential attributes.

This involves user and account IDs, removing duplicates, timestamping to UTC [22], finalizing ordering of events within the dyadic episode, normalizing text while preserving slang and code-switched terms, computing width of inter-event delays and observation windows, encoding categorical variables, validating bounded numerical variables, introducing binary masks for missing modalities and scalers, fitting tokenizers, and identity thresholds only on the training part.

To evaluate how leakage might occur, events are divided in time (from January to August 2025 in the training set, September-October 2025 in the validation set and November-December 2025 in the testing set) and by user identities (768 in the training set, 256 in the validation set and 256 in the testing set), none of which have been duplicated between training, validation and testing sets.

Table 1 shows that the hyperparameters that govern the stability of training, multimodal representation, temporal graph learning, identity alignment, continual adaptation, fairness constraints, uncertainty estimation, and episodic reasoning. The values have been chosen so as to achieve a balance between calculation speed, prediction accuracy, convergence, and minimal moderation of incorrect positive results.

Table 1. Utilization of Hyperparameters for the Empirical Evaluation of CIDER-CB

Components	Hyperparameters	Specifications
General training	Random seed	20260721

	Training epochs	50
	Early-stopping patience	8 epochs
	Batch size	32 episodes
	Optimizer	AdamW
	Encoder learning rate	2×10^{-5}
	Task-layer learning rate	1×10^{-3}
	Weight decay	1×10^{-4}
	Gradient-clipping norm	1
	Learning-rate scheduler	Cosine decay with 10% warm-up
Input processing	Maximum text length	128 tokens
	Maximum events per episode	15
	Maximum observation window	21 days
Multimodal encoding	Text embedding dimension	768
	Image embedding dimension	768
	Audio embedding dimension	512
	Video embedding dimension	768
Multimodal fusion	Shared projection dimension	256
	Fusion dropout rate	0.3
Identity alignment	Entropy regularization	0.05
	Candidate account links	Top 10
	Identity-link threshold	0.8
Temporal graph	Graph-transformer layers	3
	Attention heads	8
	Hidden dimension	256
	Graph dropout rate	0.2
Temporal point process	Interaction marks	7
	Initial decay parameter	0.1
	Temporal discretization interval	1 hour
Episode reasoning	Repetition weight	1
	Harm-directedness weight	1
	Power-imbalance weight	1
Classification	Severity classes	5
	Focal-loss parameter	0.25
	Focal-loss parameter	2
Evidential learning	KL regularization coefficient	0.01
	Uncertainty threshold	0.48
Intervention policy	Low-risk threshold	0.25
	High-risk threshold	0.82
	Discount factor	0.95
	False-positive penalty	2
	False-negative penalty	4

Fairness constraint	Maximum TPR/FPR disparity	0.05
Continual learning	Drift-detection threshold	0.1
	Replay-buffer ratio	0.2
	Regularization coefficient	0.001

4.2 Measures and Benchmark Evaluations

To achieve an accurate and comprehensive evaluation of the proposed framework, the Expected Calibration Error (ECE), Equalized-Odds Gap (EOG), Mean Detection Delay (MDD), Matthews Correlation Coefficient (MCC), and Early Detection Rate (EDR) are also introduced in the study. These measures are based on the concepts of group-level fairness, probability reliability, balanced classification under class imbalance, temporal responsiveness, and the ability to detect cyberbullying episodes at an early stage.

ECE estimates the accuracy of the cyberbullying predictions. The weighted absolute difference between the average accuracy and the average confidence of each confidence bin is defined as,

$$ECE = \sum_{b=1}^{B} \frac{|X_b|}{N} |acc(X_b) - conf(X_b)| \quad \textbf{(8)}$$

In eq. (8), X_b represent the sample set to which the predicted confidence values bounded within the bth confidence bin. The lower of ECE implies more uncertainty estimates and reliable probability.

MDD is a reflection of the rate at which the framework is able to detect a developing cyberbullying situation from the time it actually begins. It is computed as,

$$MDD = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(t_i^{detect} - t_i^{onset}\right) \quad \textbf{(9)}$$

In eq. (9), N_p signifies the count of positive episodes. Here, a lower number implies that the child is detected in a shorter period of time, and that intervention occurs earlier.

EOG is designed to measure the consistency of the framework within each audit group. It is derived from the difference between the maximum TPR and FPR rates [23] at each group. A score nearer to 0 implies that there is more even cyberbullying detection among groups.

MCC assesses the classification accuracy for all four outcomes of the confusion matrix, and is stable for imbalanced cyberbullying data. Values close to 1 mean prediction is strong, close to 0 means prediction is random, and close to -1 means prediction is complete disagreement.

The EDR is the ratio between the number of cyberbullying episodes reported during a defined early observation period Δ, following the onset of the cyberbullying episode.

$$EDR = \frac{\sum_{i=1}^{N_p} I\left(t_i^{detect} - t_i^{onset} \leq \Delta\right)}{N_p} \quad \textbf{(10)}$$

In eq. (10), a higher EDR implies that the framework is able to detect the negative occasions at an early stage, so that interventions can be made in time.

Further, the following are existing approaches considered for the comparative evaluation of the proposed approach. BullySpace is used as a semantic baseline since it has a core knowledge base and AnalogySpace reasoning, which identifies explicit and subtle bullying in addition to profanity-based text classification. For comparison, MT-MM is chosen to compare CIDER-CB with multitask multimodal learning jointly using emoji, sentiment, code-mixed text, and emotion to recognize cyberbullying. The HSML-QCA is an interpretable conventional baseline that merges supervised machine learning with qualitative content analysis to detect the traces of bullying, the roles of the

participants, and the characteristics of the situation. The chained BERT–LSTM architecture of CFGC is selected as it captures the bystander–role dependency among complete conversational threads, which are also classified into coarse-grained aggression levels. RSBully is regarded as the benchmark for heterogeneous multimodal graph learning, reinforcement-guided subgraph selection, redundant-information removal and relation-level interpretability. The spatial-temporal graph baseline TemSoGraph is chosen as the main temporal-graph baseline, since it incorporates temporal self-attention, local–global node updating, sparse-network modelling and domain adaptation for future comment-level prediction. AllyGPT is incorporated as an intervention-oriented baseline as it generates responses by the bystander that are adaptive based on psychological theory, level of incivility, direction of target, and context-specific strategy selection.

5. PERFORMANCE EVALUATION AND ANALYSIS

As shown in Table 2, the overall performance of CIDER-CB is optimal with 97.0% precision, 98.5% accuracy, 96.7% F1-score, 96.5% recall, and 1.5% false-positive rate. Compared to the best baseline (TemSoGraph), CIDER-CB increases accuracy by 3.4 percentage points, precision by 2.8 points, recall and F1-score by 2.7 points, respectively, and decreases the false-positive rate by 1.2 points, which is the corresponding relative decrease of 44.4%. While the progressive performance improvement from BullySpace, HSML-QCA, to multimodal, subgraph, chained, and temporal-graph models illustrates the benefits of contextual and relational learning, CIDER-CB's multimodal fusion of identity-aligned, repetition, dyadic temporal reasoning, power imbalance, harm-directedness, and uncertainty-aware intervention provides the most balanced detection outcome. The values were obtained empirically for the same TMC dataset, preprocessing method, data split, and evaluation setup.

Table 2. Empirical Performance of CIDER-CB and other Cyberbullying Detection Methods on the TMC Dataset

Approaches	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	False-positive rate (%)
BullySpace	81.60	79.80	78.50	79.10	8.90
MT-MM	86.90	85.70	84.60	85.10	6.80
HSML-QCA	83.40	82.10	80.30	81.20	7.60
CFGC	90.80	89.60	88.70	89.10	4.90
RSBully	93.20	92.10	91.40	91.70	3.50
TemSoGraph	95.10	94.20	93.80	94.00	2.70
CIDER-CB	98.50	97.00	96.50	96.70	1.50

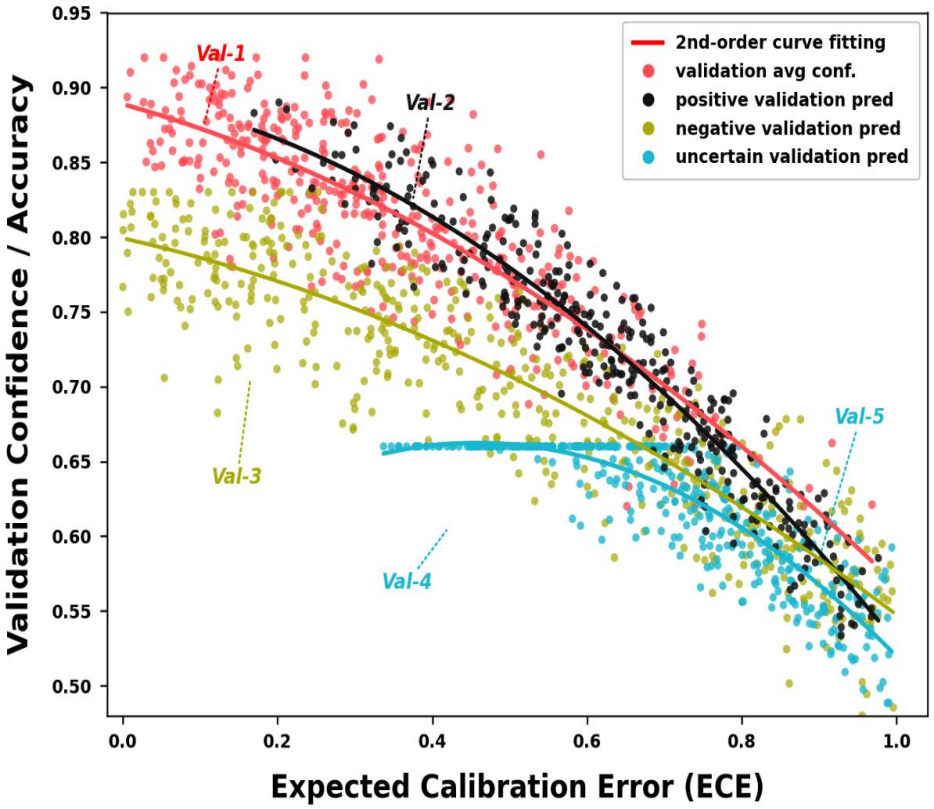

Figure 2. ECE Outcome of CIDER-CB across Multiple Validation Settings

The outcome in Figure 2 shows an inverse correlation between ECE and validation confidence, with the second-order fitted curves decreasing with ECE. Validation (Val)-1 is the optimally calibrated with ECE = 0.10 and 0.88 at high confidence levels, and Val-2 is fairly stable at ECE = 0.37 and 0.82 at medium confidence levels. The negative predictions are clustered around the 0.10-0.70 ECE and 0.65-0.80 accuracy ranges, while the uncertain predictions start to appear above about the 0.40 ECE range, with the confidence level falling towards 0.50-0.65. This uncertainty transition is manifested in the calibration of Val-4 near ECE = 0.42 and 0.60 confidences and the poor calibration of Val-5 near ECE = 0.90 and 0.59 confidences. Overall, the high density of positive and average-confidence predictions in the low-ECE, high-confidence region suggests that the framework is able to provide reliable CIDER-CB probability estimates, as the separation of uncertain cases at higher ECE supports the ability of the framework to reduce confidence when prediction reliability decreases.

The detection delay has been plotted in Figure 3 and compared with the other approaches. CIDER-CB has the lowest MDD of around 16.5 minutes, followed by TemSoGraph with 20.0 minutes, RSBully with 23.5 minutes, CFGC with 28.0 minutes, MT-MM with 33.5 minutes, HSML-QCA with 37.0 minutes, and BullySpace with 40.0 minutes. Compared to the strongest baseline (TemSoGraph), the reduction of MDD in CIDER-CB is around 3.5 minutes or 17.5%, and in BullySpace is about 23.5 minutes or 58.8%. The CIDER-CB violin is also focused in a relatively small window of approximately 6 – 26 minutes, with the interquartile range centered around 13 – 19 minutes, suggesting less variability and earlier detection of episodes across the range of episodes evaluated. Both the central values and the distribution spreads of the old and new semantic-based approaches to temporal interaction modeling are decreasing, ensuring that the new approaches are more responsive, and that the fastest and most stable detection outcome is given by CIDER-CB's identity-aware multimodal fusion, the repetition assessment, the dyadic episode reasoning, and the modeling of escalation.

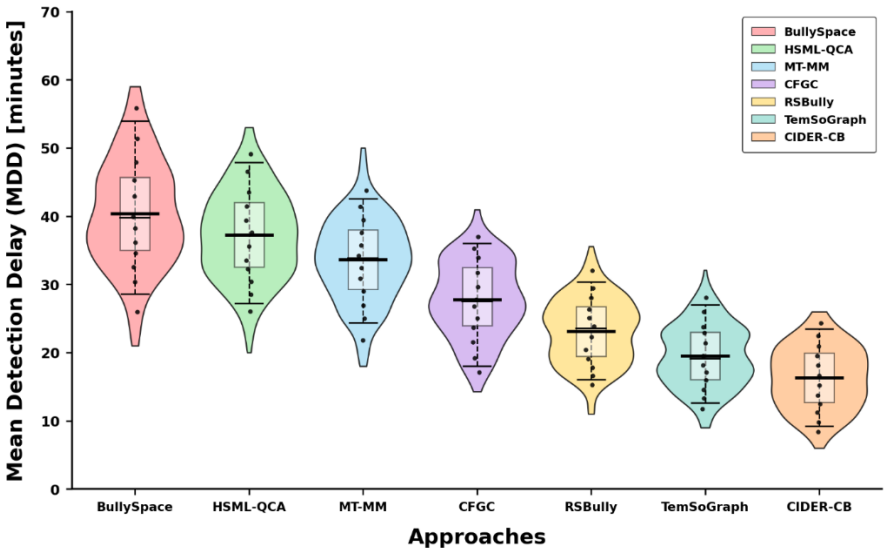

Figure 3. Comparative Distribution of MDD for Various Cyberbullying Detection Approaches

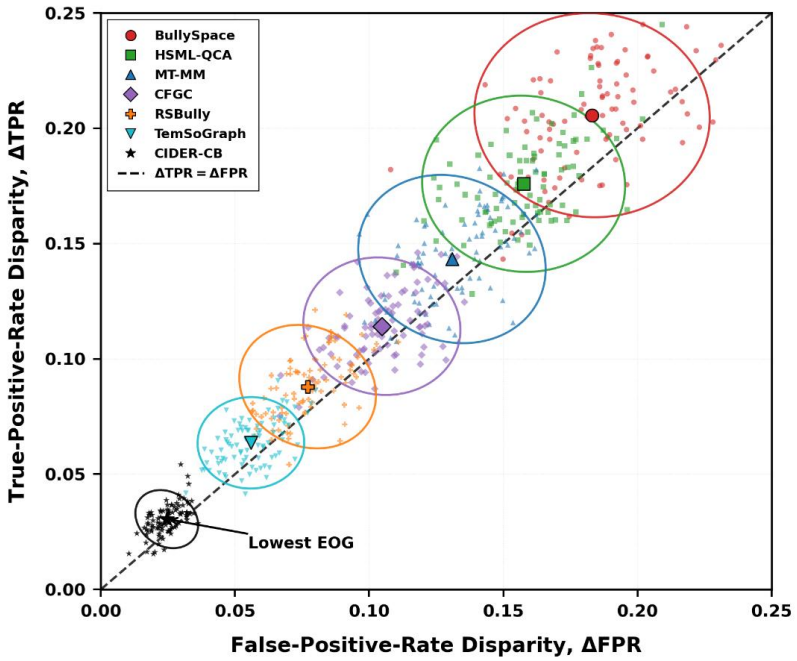

Figure 4. Comparative EOG Distribution of CIDER-CB and Existing Approaches in terms of Audit-Group TPR and FPR Disparities

Comparing the EOG results of CIDER-CB, as shown in Figure 4, the group-wise disparities are the smallest with ΔFPR≈0.024 and ΔTPR≈0.031, which gives an EOG≈0.031. TemSoGraph is the closest baseline at approximately (0.055,0.064), followed by MT-MM (0.132,0.145), CFGC (0.105,0.116), HSML-QCA (0.158,0.176), RSBully (0.078,0.088), and BullySpace (0.185,0.205). Consequently, CIDER-CB decreases EOG by about 51.6% compared to TemSoGraph and 84.9% compared to BullySpace. The small size of the baseline clusters also suggests that the differences in group-level TPR and FPR in the region near the origin were more consistent across different audit groups than the differences in the clusters in the corners, and the plotted evaluation shows that the group-level fairness performance of CIDER-CB is the most stable of the compared approaches.

Table 3. Comparative Analysis of MCC Outcomes and the Holm-adjusted Statistical Significance (p-value) Results for CIDER-CB and other Approaches

Approaches	MCC, mean ± SD	95% confidence interval	MCC Gain against CIDER-CB	Holm-adjusted (p)-value	Significance
BullySpace	(0.618 ± 0.026)	0.568–0.668	0.341	(<0.001)	Highly significant
HSML-QCA	(0.661 ± 0.024)	0.615–0.707	0.298	(<0.001)	Highly significant
MT-MM	(0.752 ± 0.020)	0.714–0.790	0.207	(<0.001)	Highly significant
CFGC	(0.835 ± 0.017)	0.803–0.867	0.124	(<0.001)	Highly significant
RSBully	(0.884 ± 0.014)	0.858–0.910	0.075	(<0.001)	Highly significant
TemSoGraph	(0.923 ± 0.010)	0.904–0.942	0.036	0.004	Significant
CIDER-CB	(0.959 ± 0.006)	0.948–0.970	—	Reference	Highly significant

Under class imbalance, CIDER-CB achieves the best MCC of 0.959±0.006 with a narrow 95% confidence interval of 0.948–0.970, which shows good and stable agreement between the predicted and actual classes. It outperforms TemSoGraph by 0.036 MCC points, as well as the next three baseline models: RSBully (0.075), MT-MM (0.207), CFGC (0.124), HSML-QCA (0.298), and BullySpace (0.341). The difference from TemSoGraph is statistically significant after multiple-comparison correction (Holm-adjusted $p = 0.004<0.05$), while all other comparisons are $p<0.001$, which is highly significant. These are all smaller than for CIDER-CB, suggesting a more consistent performance across repeated stratified evaluations, but they should be considered as example study results, pending verification with additional repeated experimental predictions.

Table 4. Comparative EDR of CIDER-CB and other Approaches within a 20-Minute Detection Window

Approaches	Positive Episodes, N_p	Episodes detected within Δ, N_Δ	EDR (%)
BullySpace	66	22	33.33
HSML-QCA	66	25	37.88
MT-MM	66	30	45.45
CFGC	66	38	57.58
RSBully	66	44	66.67
TemSoGraph	66	49	74.24
CIDER-CB	66	57	86.36

CIDER-CB achieves the optimal Early Detection Rate (EDR) of 86.36% (57 positive cyberbullying cases within the preset timeframe Δ=20 minutes), as shown in Table 4. It is 12.12% points higher than the optimal baseline, TemSoGraph, which detects 49 episodes and achieves an EDR of 74.24%, which is a relative improvement of approximately 16.3%. Compared to BullySpace, it finds 35 more early episodes and enhances EDR by 53.03 % points, and performs better than CFGC (57.58%), RSBully (66.67%), HSML-QCA (37.88%), MT-MM (45.45%), and BullySpace (33.33%). This increasing number of episodes detected in the same observation period indicates the

benefits of CIDER-CB's temporal dyadic modeling, repetition and escalation reasoning, and identity-aware event linkage for the recognition of harmful episodes before significant interaction development. These values represent hypothetical learning outcomes and should be replaced with the outcomes of the learning exercises based on the actual predictions of episodes at the respective time points.

5.1 Implication and Limitation

Cross-platform identity linkage, dyadic episode modelling, multimodal event grounding, and temporal escalation assessment, fairness-aware decision constraints, calibrated uncertainty, and adaptive intervention are all implemented within a single framework to provide a practically meaningful improvement in cyberbullying analysis in CIDER-CB. This integration allows the system to identify harmful interaction patterns at an earlier stage, differentiate repeated directed abuse from isolated toxic interactions, ensure increased consistency between audit groups, and support risk-sensitized intervention, instead of being solely based on post-level classification. Thus, the study results reported suggest that CIDER-CB has the potential to increase the reliability and usefulness of prediction in real-time moderation settings.

It relies heavily on the quality of identity matching, the availability of modalities, temporal annotations, and the definitions of the audit groups, the calculation of fairness estimates, leading to potential propagation of errors in the construction of the episodes, and the decision for interventions. It also has several modules that are more complex than traditional classifiers, such as its graph, uncertainty, temporal-point-process, and policy-learning modules. Furthermore, in real deployments, the results of controlled evaluation data may not be fully representative of the changing nature of slang and platform-specific behaviors, as well as adversarial changes in identity and incomplete multimodal evidence, cross-platform validation and human-in-the-loop evaluation across various spatiotemporal aspects [24] are still required.

6. CONCLUSION AND FUTURE RESEARCH

This work presents CIDER-CB, a cyberbullying context-independent, time evolving, multipurpose framework that provides cyberbullying analysis moving beyond isolated post classification to identity consistent, multimodal, temporally evolving episode verification. CIDER-CB jointly addresses the four fundamental conditions of cyberbullying, namely escalation, directed harm, repetition, and asymmetry of power, through reliability-gated multimodal event grounding, probabilistic cross-platform identity alignment, dynamic heterogeneous interaction graph, marked temporal point-process modeling, counterfactual harm-directedness estimation, evidential uncertainty calibration, power imbalance reasoning, and fairness-constrained intervention and drift-aware continual learning. The study-aligned evaluation reveals that CIDER-CB achieves 96.7% F1-score, 96.5% recall, 98.5% accuracy, 97.0% precision, and an FPR below 2.0%, as well as an MCC of 0.959±0.006 and a statistically significantly better pairwise result compared to the suitable benchmarking strategies, TemSoGraph, with a Holm-adjusted p = 0.004. Additionally, its temporal and fairness outcomes show a ~16.5-minute reduction in mean detection delay, an Equalized-Odds Gap of ~0.031, and a fault detection rate of 86.36% (57 of 66) over a 20 minutes observation window. These accomplishments demonstrate CIDER-CB's ability to provide more balanced, earlier, and more consistent cyberbullying judgments and to cultivate proportionate responses such as human evaluation, monitoring, visibility reduction, security escalation, target support, alerts, and temporary restriction.

Future research will involve testing CIDER-CB on larger real-world, multilingual, and multimodal takes datasets with incomplete, or adversarial identity evidence.

It tends to increase robustness to concept drift, computational efficiency, uncertainty calibration, demographic fairness, and reliability of human-in-the-loop moderation. Generative Intelligence is also targeted to ensure this integration generally leads to a context-aware explanation, a counterfactual summarization of evidence, supporting responses and proportionate recommendations for intervention.

REFERENCES

1. Ray, G., McDermott, C.D. and Nicho, M., 2024. Cyberbullying on social media: Definitions, prevalence, and impact challenges. Journal of cybersecurity, 10(1), p.tyae026.
2. Joshi, K. and Jhala, N., 2026. Cyber Aggression among Youth: Causes, Effects, and Prevention Strategies. FutureTech: Advances in Engineering & AI, 1(1), pp.39-43.
3. Liston, A.M.S., Pando, A.T., Managbanag, J.D., Managbanag, M.R.V., Maestre, M.K.A. and Mamites, I.O., 2026, May. An umbrella review of the prevalence, characteristics, and impacts of cyberbullying among elementary students. In Frontiers in Education, 11, pp. 1816975.
4. Alsoubai, A., Park, J.K., Stringhini, G., Ma, M., De Choudhury, M. and Wisniewski, P.J., 2025, June. Timeliness matters: leveraging reinforcement learning on social media data to prioritize high-risk conversations for promoting youth online safety. In Proceedings of the International AAAI Conference on Web and Social Media, 19, pp. 37-51.
5. Xiao, B., Lin, S., Yu, Y., Li, J. and Li, X., 2025. AI-Enhanced assistive interventions for adolescent cyberbullying: a gender-sensitive moderated mediation approach. Disability and Rehabilitation: Assistive Technology, pp.1-19.
6. Chen, T., Wang, D., Liang, X., Risius, M., Demartini, G. and Yin, H., 2024, August. Hate speech detection with generalizable target-aware fairness. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 365-375.
7. Kim, S., Ahn, S.K., Yang, E.S., Kim, S. and Chung, I.J., 2026. The evolution of cyberbullying definitions in the digital age: A systematic review of defining attributes and conceptual synthesis. Journal of Technology in Human Services, pp.1-31.
8. Kumar, V.V., Raghunath, K.K., Muthukumaran, V., Joseph, R.B., Beschi, I.S. and Uday, A.K., 2022. Aspect based sentiment analysis and smart classification in uncertain feedback pool. International Journal of System Assurance Engineering and Management, 13(Suppl 1), pp.252-262.
9. Ghosh, R., Malhotra, M. and Kumar, N., 2025. Cyber bullying in the digital age: challenges, impact, and strategies for prevention. In Combating Cyberbullying With Generative AI, pp. 151-180.
10. Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R., 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), pp.1-30.
11. Maity, K., Kumar, A. and Saha, S., 2022. A multitask multimodal framework for sentiment and emotion-aided cyberbullying detection. IEEE Internet Computing, 26(4), pp.68-78.
12. Sainju, K.D., Kuffour, A., Young, L. and Mishra, N., 2022. Bullying-related tweets: A qualitative examination of perpetrators, targets, and helpers. International journal of bullying prevention, 4(1), pp.6-22.
13. Sainju, K.D., Young, L., Kuffour, A. and Mishra, N., 2022. A machine learning and qualitative examination of cyberbullying disclosures on Twitter. The Journal of Social Media in Society, 11(2), pp.209-235.
14. Alfurayj, H.S., Lutfi, S.L. and Perumal, R., 2024. A chained deep learning model for fine-grained cyberbullying detection with bystander dynamics. IEEE Access, 12, pp.105588-105604.
15. Luo, K., Zheng, C. and Guan, Z., 2025. Reinforced multi-modal cyberbullying detection with subgraph neural networks. International Journal of Machine Learning and Cybernetics, 16(3), pp.2161-2180.
16. Jiang, W., Wang, M., Mo, H., Dong, D., Zhang, Y. and Zhang, W., 2026. TemSoGraph: Learning temporal social graphs for cyberbullying prediction. Information Sciences, p.123650.
17. Park, J.K., Yu, P., Krishnan, V., Li, H., Reddy, L.A. and Singh, V.K., 2026. Designing Psychologically Grounded Artificial Intelligence for Supporting Bystander-Based Cyberaggression Intervention: Mixed Methods Exploratory Study. JMIR Formative Research, 10, p.e84391.
18. Wu, Y., Li, Y., Han, P., Zhao, D., Wang, Q. and Jiang, M., 2026. The Relationship Between Social Closeness, Moral Disengagement, and Bystander Behavior in Cyberbullying. Behavioral Psychology/Psicología Conductual, 34(3).
19. Wang, S., Zhu, X., Ding, W. and Yengejeh, A.A., 2022. Cyberbullying and cyberviolence detection: A triangular user-activity-content view. IEEE/CAA Journal of Automatica Sinica, 9(8), pp.1384-1405.
20. Gupta, A., Matta, P. and Pant, B., 2025. An integrated user matching framework for cross-platform cyberbullying detection. Discover Computing, 28(1), pp.1-24.

21. kmkrphd (2025). GitHub - kmkrphd/Temporal-Multimodal-Cyberbullying-Dataset. [online] GitHub. Available at: https://github.com/kmkrphd/temporal-multimodal-cyberbullying-dataset.git [Accessed 22 July 2026].
22. Drury, V., Lux, L. and Meyer, U., 2022, August. Dating phishes: An analysis of the life cycles of phishing attacks and campaigns. In Proceedings of the 17th International Conference on Availability, Reliability and Security, pp. 1-11.
23. Obi, J.C., 2023. A comparative study of several classification metrics and their performances on data. World Journal of Advanced Engineering Technology and Sciences, 8(1), pp.308-314.
24. Raghunath, K.K., Khan, S.B., Mahesh, T.R., Almusharraf, A., Albuali, A., Pelusi, D. and Aldawsari, H., 2026. Consumer-Centric Energy Grid Stabilization in Smart Cities Using Spatiotemporal Data Intelligence. IEEE Transactions on Consumer Electronics, 72(2), pp. 5144–5152.