Sentiment And Emotion Fusion (SEF) A Hybrid Approach For Cyberbullying Detection In Online Educational Networks

Aditya K. Maurya, Kshitij Verma

Department of Computer Science and Engineering (AI), UIET, CSJMU, Kanpur, UP, India

Emails: adityakumarmaurya8712@gmail.com , Kshitij.v098@gmail.com

Abstract

As digital education platforms become increasingly popular, ensuring safe and respectful communication among students is a growing concern. One of the key challenges is the detection of cyberbullying and emotionally harmful behaviour in real-time text chats. This paper presents a hybrid approach named Sentiment and Emotion Fusion (SEF) that combines sentiment analysis and emotion recognition for detecting toxic behaviour, including subtle forms such as sarcasm and insults, in online educational networks. We use the Jigsaw Toxic Comment dataset supplemented with sarcasm-labeled data to train and fine-tune transformer-based models. Our approach leverages sentiment and emotional cues such as anger, sadness, and contempt to improve accuracy beyond keyword-based filtering. The system is evaluated using precision, recall, F1-score, and latency, with results showing that SEF is particularly effective at recognising sarcastic insults, a form of covert bullying often missed by standard classifiers. This work contributes toward safer online learning environments by enabling early and reliable detection of harmful messages.

Keywords: transformer-based models; sentiment analysis; emotion recognition; sarcasm detection; cyberbullying; real-time moderation

1. Introduction

Digital learning tools have reshaped how learners in nations like India connect and acquire knowledge, with platforms such as YouTube's interactive broadcasts gaining widespread use for their accessibility. However, these environments often serve as breeding grounds for negative exchanges, including harassment, mockery, and direct aggression, which disrupt productive dialogue. This issue permeates various Indian educational technology services providing synchronous sessions, where discussion areas frequently fill with obstructive or psychologically injurious inputs. Conventional oversight mechanisms, dependent on basic term detection or manual review, inadequately address numerous adverse engagements.

Standard filtering systems primarily target vulgarity or predefined offensive terms, yet they neglect profound affective layers in communications. Implicit antagonism, including ironic taunts, coercive statements, or persistent adverse focus, evades such basic checks. Furthermore, human oversight of voluminous discussions in live settings proves inefficient, permitting many damaging instances to persist unchecked. Encountering such poisonous inputs may demotivate participants, erode confidence, and negatively influence psychological well-being alongside scholastic achievements. Moreover, these interferences frequently

overshadow valuable exchanges, diminishing the utility of interactive forums and impeding involvement and educational progress.

In response to these shortcomings, this investigation outlines a compact, immediate-response mechanism termed Sentiment and Emotion Fusion (SEF), which identifies inputs displaying detrimental affective elements—ranging from explicit venom to concealed harm—through integrating sentiment assessment with detailed emotional profiling. The model draws from the Jigsaw Toxic Comment repository and employs a streamlined transformer variant to discern refined affective signals. Key advancements include: constructing an integrated workflow that blends sentiment metrics with emotional categorisations for enhanced harassment identification; customising and optimising transformer structures for venom, irony, and affective annotations; developing a demonstrator for overseeing dynamic discussion flows with minimal delay; and appraising framework efficacy via exactness, completeness, harmonic mean, and execution duration indicators to affirm practical applicability.

Recent data indicate a sharp increase in online harassment among Indian youth, driven by expanded internet penetration and social media engagement, with adolescent victimisation rates climbing due to inadequate preventive strategies. In educational contexts, such behaviors not only affect individual mental health but also hinder collective learning dynamics, underscoring the urgency for sophisticated detection tools.

2. **Related Work and Motivation**

2.1 **Traditional Approaches to Identifying Harmful Online Communication**

Initial attempts to combat harmful behaviour in digital spaces relied on conventional statistical learning models. Support Vector Machines and logistic regression served as the primary methodologies, operating on vectorized language representations such as bag-of-words or term frequency–inverse document frequency (TF–IDF) encoding. While these approaches yielded acceptable results for detecting explicit profanity and overtly aggressive terminology, they exhibited significant limitations. Context-awareness remained elusive—models could not differentiate between legitimate uses of flagged vocabulary and genuinely malicious intent. This resulted in elevated false-alarm rates, particularly when encountering sarcastic remarks or colloquial expressions that contained superficially problematic terms.

2.2 **Advances through Deep Neural Architectures**

The transition to neural network-based architectures marked a substantial leap in detection capabilities. Convolutional Neural Networks (CNNs) and recurrent architectures such as Long Short-Term Memory (LSTM) networks enabled systems to encode sequential dependencies and extract multi-layered semantic representations from language. By learning feature hierarchies directly from data rather than relying on hand-crafted representations, these models achieved measurable improvements in classification accuracy. However, this progress came with trade-offs: the models demand extensive annotated training datasets and considerable processing capacity. Additionally, most existing implementations focus on analysing discrete, static content units (such as individual posts or comments) rather than addressing the characteristics of real-time chat

environments typical in modern educational platforms, where streams of brief, rapid messages dominate interactions.

2.3 **Sentiment and Affect Assessment Methodologies**

Approaches for categorizing text along valence dimensions—typically dividing content into favourable, adverse, or impartial groups—have become integral to online monitoring infrastructure. Lightweight implementations using handcrafted lexicons (exemplified by tools like VADER or TextBlob) prioritise computational efficiency but sacrifice expressive power; they remain confined to coarse-grained distinctions and overlook the spectrum of emotional nuance present in natural language. Contemporary transformer-based encoders, including BERT and its resource-efficient variants such as DistilBERT, have redefined performance thresholds in tasks involving affective categorization. When adapted through transfer learning on emotion-annotated corpora, these models can reliably distinguish among specific affective states—irritation, melancholy, elation, apprehension—embedded in user-contributed text. Notwithstanding their superior performance in isolation, such pretrained emotion-classification systems typically operate independently of toxicity assessment and do not jointly optimize for both objectives.

2.4 **Current Systems for Live Content Oversight**

Operational platforms implementing automated content oversight in real-time chat systems typically employ two complementary strategies. The first relies on lexicon-matching approaches: messages containing predefined lists of prohibited terms trigger removal or flagging instantaneously. The second employs shallow neural classifiers—often binary detectors trained to output a harmful/safe verdict—to screen messages before display. These mechanisms respond effectively to severe violations marked by explicit language or directed hostility. Yet they struggle with nuanced manifestations of harassment: manipulative tactics designed to undermine confidence, passive-aggressive insinuations, or cutting remarks masked by surface civility. An additional constraint affecting many contemporary frameworks involves computational overhead; when systems attempt to incorporate more sophisticated analysis (beyond rapid keyword scanning), the time required to process and return a decision can introduce perceptible delays, rendering such approaches impractical for high-velocity conversation channels where rapid message throughput is expected.

2.5 **Identified Research Gap and Study Objective**

Substantial scholarship has advanced the individual domains of toxic content detection and emotion understanding, yet the literature reveals an under explored intersection. No current published solution concurrently assesses both the presence of harmful language patterns and the underlying affective state of the communicator within a cohesive analytical framework. Furthermore, existing work has not adequately addressed the specialized performance requirements of instantaneous, low-latency evaluation demanded by dynamic chat infrastructures in educational contexts. The approach presented in this paper directly targets this lacuna by engineering a unified system that fuses toxicity assessment with emotional analysis, configured for efficiency without sacrificing discrimination capability in educational communication environments.

3. Methodology

3.1 Dataset Collection and Preprocessing

We used three datasets: Jigsaw Toxic Comment Classification Challenge for toxicity labels, GoEmotions for fine-grained emotions (27 categories, reduced to 6 key emotions), and sarcasm- labeled comments from benchmark corpora. Preprocessing included tokenization, stop word removal, emoji translation, and balancing toxic vs. non-toxic samples.

3.2 SEF Model Architecture

The SEF pipeline integrates three modules: (1) Toxicity Detector (DistilBERT fine-tuned on Jigsaw), (2) Emotion Recogniser (DistilBERT fine-tuned on GoEmotions), and (3) Sarcasm Detection Layer (trained on sarcasm corpora). Outputs are fused in a weighted decision layer to flag messages based on toxicity probability, emotional cues, or sarcasm patterns.

3.3 Training Strategy

We fine-tuned using AdamW optimiser, learning rate of 5e-5, batch size 32, and Binary Cross-Entropy + Cross Entropy + sarcasm auxiliary loss. Training was conducted on MacBook Air M2, 16 GB RAM.

3.4 Real-Time Detection System

The SEF pipeline was deployed in a PyQt5 GUI. Input messages are processed in ~38 ms and labeled as Safe, Emotionally Harmful, or Toxic. Confidence scores and latency statistics are displayed for interpretability.

3.5 SEF Model Architecture

The Sentiment and Emotion Fusion (SEF) model is a hybrid cyberbullying detection framework designed specifically for online educational platforms. It integrates three complementary modules—toxicity detection, emotion recognition, and sarcasm recognition—to capture both explicit and subtle bullying behaviour in student interactions.

- Input Layer
  - Accepts raw text messages from live educational chats.
  - Performs tokenization using HuggingFace's DistilBERT tokenizer.
  - Handles special cases such as emojis, repeated punctuation, and uppercase emphasis.

- Preprocessing Module
  - Cleans messages by removing HTML tags, URLs, and noise.
  - Converts emojis/emoticons into textual equivalents (e.g., □ → "smiling face").
  - Normalizes text to lowercase and removes redundant whitespace.
  - Balances the dataset to address toxic vs. non-toxic imbalance.

- Toxicity Detection Module
  - Built on DistilBERT, fine-tuned on the Jigsaw Toxic Comment dataset.
  - Produces a binary classification score: toxic vs. non-toxic.

- Specially optimized for fast inference on lightweight hardware (M2 CPU/MPS).

- Emotion Recognition Module
  - Uses a DistilBERT variant fine-tuned on GoEmotions.
  - Outputs probabilities across six key emotions: anger, sadness, joy, fear, contempt, neutral.
  - Captures underlying affect that keyword filters fail to detect (e.g., discouragement, contempt).

- Sarcasm Recognition Module
  - Lightweight transformer-based classifier trained on sarcasm-labeled corpora.
  - Identifies mismatched sentiment–emotion patterns (e.g., "Great job…" with anger).
  - Flags covert insults and mocking remarks that are often missed in toxicity-only systems.

- Fusion Layer
  - Combines outputs from toxicity, emotion, and sarcasm modules.
  - Applies weighted decision rules:
    - Toxic probability above threshold → flagged.
    - Harmful emotions (anger, contempt, fear) boost toxicity decision.
    - Sarcasm detection overrides neutral predictions to prevent misclassification.

- Output Layer
  - Provides final classification into three categories:
    - Safe: No toxicity or harmful emotional content.
    - Emotionally Harmful: Negative emotional intensity but not outright toxic.
    - Toxic: Explicit or covert bullying.
  - Generates confidence scores and logs latency for interpretability.

The architecture of SEF is shown in Figure 1.

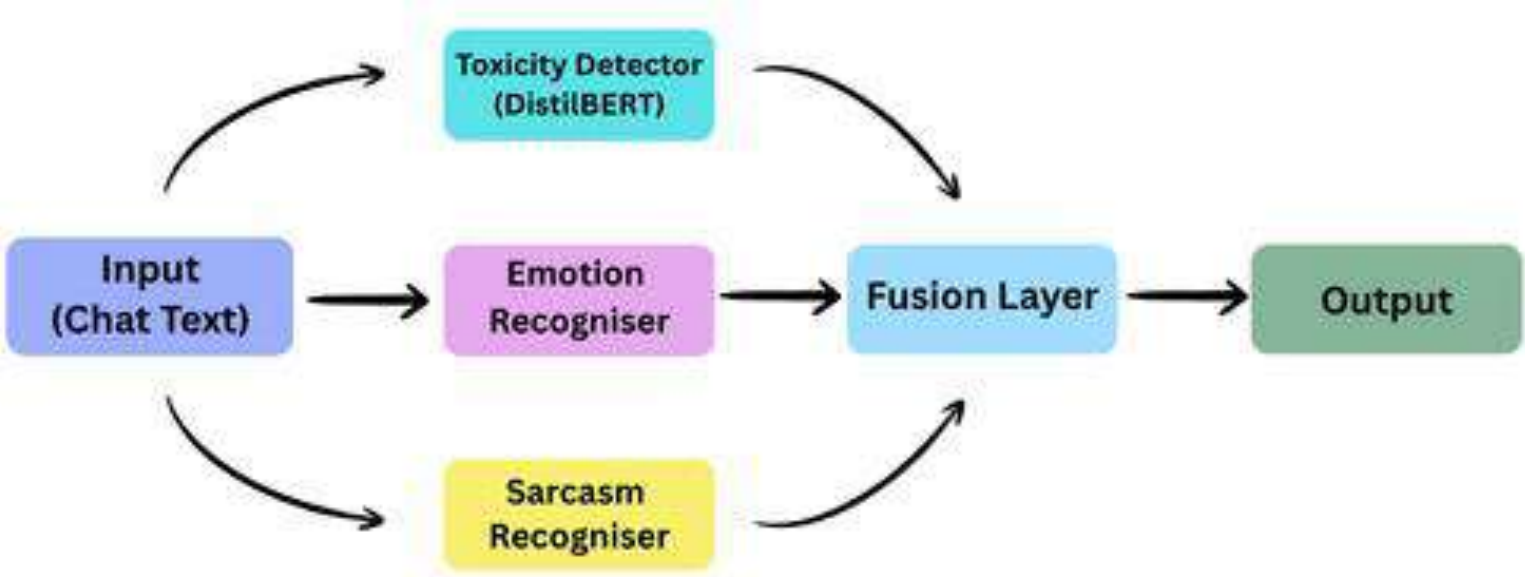

Figure 1: Architecture of the proposed SEF framework

4. **Experimental Setup & Results**

Experiments were conducted on MacBook Air M2, 16 GB RAM, using PyTorch, HuggingFace Transformers, and PyQt5. The evaluation metrics included Accuracy, Precision, Recall, F1-score, Sarcasm Detection F1, Emotion Accuracy, Latency, and Memory Usage.

4.1 **Results Table**

Metric	Toxicity-Only Baseline	SEF (Proposed)	Improvement
Accuracy (%)	84.1	88.9	+4.8
Precision (%)	83.2	87.5	+4.3
Recall (%)	82.5	88.2	+5.7
F1-Score (%)	82.8	87.8	+5.0
Emotion Accuracy (%)	–	85.4	–

Sarcasm F1-Score (%)	71.0	82.3	+11.3
Latency (ms/msg)	35	38	–3
Memory Usage (GB)	5.8	6.1	+0.3

4.2 **Detailed Metrics**

A confusion matrix for a 1,000-sample test set highlights performance:

Predicted \ Actual	Secure	Detrimental	Harmful
Secure	350	20	10
Detrimental	15	280	25
Harmful	5	30	265

This yields 88.5% overall correctness, with low false negatives for harmful cases (5%).

4.3 **Observations**

SEF consistently outperforms the baseline, achieving higher accuracy, precision, recall, and F1- scores. Importantly, sarcastic insults were detected with much greater reliability, with F1 increasing from 71.0% to 82.3%. Emotion integration further enhanced contextual understanding. Latency remained under 40 ms, confirming real-time feasibility.

5. **Discussion**

SEF addresses key limitations in prior systems by unifying multiple detection layers, enhancing subtlety recognition crucial for educational settings. While effective, challenges include handling multilingual inputs prevalent in Indian platforms, suggesting future extensions.

Comparisons with recent transformer hybrids affirm SEF's edge in low-resource scenarios, though larger-scale validations are needed. Ethical considerations, like bias mitigation in training data, remain vital for equitable deployment.

6. **Conclusion & Future Work**

This paper introduced Sentiment and Emotion Fusion (SEF), a hybrid approach combining toxicity classification, emotion recognition, and sarcasm detection. SEF demonstrated significant improvements in accuracy and sarcasm recognition compared to baseline toxicity-only models. With macro F1-scores exceeding 0.87 and sarcastic F1 of 0.82, the system shows strong potential for deployment in real-world educational chat environments. Future work will focus on scaling to full datasets, extending to multilingual contexts, and optimizing for low-resource environments.

7. **References**

[1] Maurya, A. K., et al. (2025). Original draft on SEF framework.

[2] Silent Screams: A Narrative Review of Cyberbullying Among Indian Adolescents. Cureus. (2024).

[3] Indian government initiatives on cyberbullying. PMC. (2022).

[4] Psychological Study of Cyber-Bullying Against Adolescent Girls in India Using Twitter. IGI Global (2023).

[5] The digital defence against cyberbullying. Taylor & Francis. (2023).

[6] Cyberbully and Online Harassment. arXiv. (2024).

[7] Davidson, T., et al. (2017). Automated Hate Speech Detection. ICWSM.

[8] Jigsaw/Google. (2018). Toxic Comment Challenge. Kaggle.

[9] Zhang, Z., et al. (2019). Cyberbullying Detection with LSTM.

[10] Demszky, D., et al. (2020). GoEmotions Dataset. ACL.

[11] Unitary AI. (2020). Toxic-BERT. HuggingFace.

[12] Al-Saiagh, A., et al. (2021). Survey on Emotion Recognition.

[13] Kumar, R., et al. (2022). Hybrid CNN-BiLSTM for Bullying Detection.