EmpathAI: An Intelligent Empathetic Assistant for Harassment Detection and Supportive Response Generation

1st Tharun P
Dept. of Computer Science and Business Systems
NMIT Bengaluru
Bengaluru, India
1nt23cb059.tharun@nmit.ac.in

2nd Sagar M
Dept. of Computer Science and Business Systems
NMIT Bengaluru
Bengaluru, India
1nt23cb048.sagar@nmit.ac.in

3rd Hariprasad S
Dept. of Computer Science and Business Systems
NMIT Bengaluru
Bengaluru, India
1nt23cb017.hariprasad@nmit.ac.in

***Abstract*—Online communication has increasingly become vulnerable to cyberbullying and digital harassment, both of which pose significant risks to emotional and psychological well-being. Existing automated systems often offer fragmented support, focusing either on detecting harmful language or analyzing emotional tone, but rarely combining the two into a unified and meaningful framework. This paper presents EmpathAI, a web-based empathetic assistant designed to integrate emotional intelligence, harassment detection, and legal reasoning into a single, cohesive system. EmpathAI employs fine-tuned transformer-based models—DistilRoBERTa for emotion recognition and BERT for harassment and toxicity classification—enhanced through hybrid keyword analysis for improved contextual understanding. Empathetic and context-aware responses are generated using the Google Gemini API, supported by a rule-based fallback mechanism to ensure reliability. A dedicated legal reasoning module provides localized guidance by linking severe harassment cases to relevant sections of the Indian Penal Code (IPC), bridging the gap between emotional support and legal literacy. The frontend, developed using ReactJS, preserves user privacy through local data storage and integrates a proactive Notification Monitor that identifies distressing content in real time. Through this multidisciplinary approach, EmpathAI represents a step toward human-centered, ethically grounded, and privacy-conscious artificial intelligence that not only detects online harm but also responds with compassion, awareness, and actionable guidance.**

***Index Terms*—Cyberbullying Detection, Emotion Recognition, Empathy-Based Artificial Intelligence, Ethical Computing, Gemini API, Indian Penal Code, Large Language Models, Legal Reasoning, Privacy-Preserving AI, Transformer Models.**

I. Introduction

Online social platforms have become central to communication, collaboration, and self-expression, but they also expose users to cyberbullying and digital harassment, which can seriously affect mental health and well-being [1, 2]. While modern moderation pipelines can flag toxic or abusive content, most systems still treat the problem as a narrow text classification task, offering labels rather than holistic support to affected individuals [4, 15]. Users are typically informed that content is "toxic" or "abusive," but receive little emotional validation or guidance on what to do next.

Recent advances in transformer-based language models and large language models (LLMs) have enabled richer understanding of emotional context and more human-like dialogue, opening the possibility of systems that not only detect harm but also respond with empathy [3, 7, 8]. At the same time, the integration of explainable NLP, legal reasoning, and privacy-preserving architectures has highlighted the need for AI systems that are not only accurate but also transparent, safe, and aligned with user rights [5, 6, 11, 17, 20].

EmpathAI is designed in this context as a human-centered, web-based assistant that unifies emotion recognition, harassment detection, empathetic response generation, and basic legal awareness into a single framework. Instead of acting solely as a content filter, EmpathAI aims to function as an empathetic companion that can understand users' emotional states, identify potential harassment, highlight relevant Indian Penal Code (IPC) provisions when appropriate, and respond in a supportive and privacy-conscious manner.

A. Motivation

Most existing tools for handling online harm focus on detecting and removing abusive content, often using keyword filters or toxicity scores. These approaches rarely consider how victims feel, what kind of support they might need, or whether the incident may have legal implications [2, 15]. As a result, individuals facing harassment may still feel isolated, unsafe, and unsure about their rights and options.

In countries like India, where legal protections exist but are not widely understood, victims may struggle to recognize when online behaviour crosses into legally actionable harassment under the IPC or IT Act [9, 12, 16]. At the same time, growing awareness of data privacy and AI ethics means that any supportive assistant must minimize data collection and give users control over their information [6, 18].

These gaps motivate the development of EmpathAI as a system that: (i) detects harmful communication, (ii) interprets emotional context, (iii) offers empathetic, safety-aware responses, and (iv) introduces accessible legal awareness, all

while preserving user privacy. The overarching goal is to shift from reactive moderation to proactive, human-centered support.

B. Objectives

The main objective of this work is to design and implement an intelligent, privacy-aware assistant that integrates emotion understanding, harassment detection, empathetic dialogue, and legal reasoning in a unified architecture. Specifically, EmpathAI pursues the following objectives:

- O1: Emotion-Aware Understanding — Develop a transformer-based emotion recognition module capable of reliably identifying user emotions such as anger, sadness, fear, and joy in real time [7, 19].
- O2: Robust Harassment Detection — Implement a hybrid harassment detection component that combines contextual transformer models with keyword-enhanced rules to capture explicit, implicit, and multilingual abuse with high recall [4, 8, 15].
- O3: Empathetic and Safe Response Generation — Integrate an LLM-based response generator (Google Gemini) with safety prompts and a rule-based fallback mechanism to produce empathetic, supportive, and non-harmful responses tailored to emotion and severity [3, 10, 18].
- O4: Legal Awareness and Civic Support — Design a lightweight legal reasoning module that links high-severity harassment to relevant IPC sections and explains them in accessible language, without replacing professional legal advice [9, 12, 16].
- O5: Privacy-Preserving, Real-Time Architecture — Realize a web-based, privacy-first architecture in which sensitive data remains on-device, while the backend performs anonymized inference with low latency, aligned with responsible and trustworthy AI practices [6, 11, 17].

Together, these objectives guide the design and implementation of EmpathAI as a practical, ethically grounded system that responds to the technical, emotional, and legal dimensions of online harassment in an integrated manner.

II. Literature Review

Research in emotion recognition, harassment detection, conversational AI, and ethical computing provides the conceptual backbone for the development of EmpathAI. Recent surveys in affective computing highlight the shift from shallow models to deep neural architectures for modeling human emotions in text, audio, and multimodal signals [13, 19]. Over the years, artificial intelligence has evolved from simple text classifiers into sophisticated systems capable of understanding emotional tone, identifying harmful communication, and responding with contextually appropriate empathy. Despite this progress, most existing systems remain fragmented, focusing on individual components such as toxicity detection or emotion analysis, without integrating these dimensions into a unified, human-centered framework that also considers ethical and legal reasoning.

Modern approaches rely on contextual embeddings and transformer architectures that capture long-range semantic dependencies more effectively than traditional word embeddings [7, 14]. DistilRoBERTa, a compressed variant of RoBERTa, preserves much of the performance of the full model at lower computational cost, making it highly suitable for real-time inference scenarios such as supportive chatbots [7, 19]. These advancements have established transformers as the state-of-the-art in emotion analysis, enabling systems like EmpathAI to deliver accurate, context-aware emotion recognition with reduced latency.

In parallel, the field of harassment and toxicity detection has progressed rapidly in response to the growing prevalence of online abuse. Recent studies on abusive language and cyberbullying detection emphasize the importance of contextual modeling and multilingual robustness, demonstrating the superiority of transformer-based models over keyword-only methods [4, 8, 15]. Hybrid systems that combine contextual embeddings with rule-based or keyword-enhanced layers have been shown to improve robustness on noisy social media text [2, 15]. Severity-based categorization—classifying harassment along a graded scale (low, medium, high)—has emerged as a practical and ethically grounded strategy, as it supports proportional automated actions and nuanced user feedback [2].

The rise of conversational AI has significantly expanded the potential for empathetic and emotionally intelligent systems. Recent work explores LLM-driven empathy generation, demonstrating that large language models, when guided with appropriate prompts and safety constraints, can produce emotionally supportive responses suitable for mental health and well-being contexts [3, 10]. Empathetic dialogue systems increasingly integrate emotion-aware conditioning, user history, and context preservation to improve perceived authenticity and user trust. However, such models are still vulnerable to hallucinations, inconsistent tone, and occasional unsafe outputs, motivating the development of hybrid frameworks with rule-based fallback mechanisms and safety layers [10, 18].

Legal and ethical reasoning in AI systems has also received growing attention. Legal NLP approaches are being developed to support automated extraction, retrieval, and explanation of legal provisions relevant to online harms, consumer rights, and digital safety [9, 16]. Within the Indian legal system, provisions in the IPC and IT Act are increasingly being studied in the context of online harassment and cybercrime [12]. At the same time, broader AI ethics and governance literature emphasizes accountability, transparency, fairness, and user autonomy as core design principles for human-centered AI systems [6, 11, 17, 20]. Privacy-preserving conversational agents that minimize data collection and maximize on-device processing are especially relevant for emotionally sensitive domains [18].

Collectively, existing literature demonstrates substantial advancements across emotion recognition, toxicity detection, empathetic AI interaction, and responsible AI design. Transformer-based models dominate both classification and

generation tasks, achieving high interpretive precision and contextual awareness [7, 8]. LLM-driven systems have expanded the potential for empathetic responses, while privacy-centric architectures and legal reasoning modules have redefined AI's role in fostering digital well-being [3, 9, 16]. Nevertheless, current systems often operate in isolation—detecting emotions or toxicity without offering comprehensive, actionable support.

EmpathAI seeks to bridge this gap by unifying these dimensions within a single conversational ecosystem. Through its integration of transformer-based emotion and harassment analysis, LLM-driven empathy generation, and contextually grounded legal reasoning, the system exemplifies a new paradigm of responsible and human-centered AI. In doing so, it transforms digital assistance from reactive moderation into proactive empathy and advocacy, positioning artificial intelligence not merely as a detector of harm but as a companion capable of understanding, guiding, and supporting users in moments of vulnerability.

III. Methodology

The development of EmpathAI followed a structured yet human-centered methodology. Rather than optimising only for model accuracy, each design choice was made with the user's emotional experience, safety, and privacy in mind. This section walks through how the system is organised, how the models were trained, and how all components work together in practice.

A. System Framework and Data Flow

EmpathAI adopts a two-tier modular framework consisting of a lightweight frontend client and a high-performance backend inference engine. The frontend is implemented using ReactJS and provides a soft-toned chat interface designed to feel calm and non-judgmental. All chat histories are stored locally using the browser's `localStorage` API so that sensitive messages never leave the user's device.

The backend, implemented in Python using FastAPI, orchestrates the complete inference pipeline. Incoming messages are transmitted over HTTPS or WebSocket channels and pass through four logical stages: emotion recognition, harassment analysis, legal reasoning, and empathetic response generation. This clean separation between presentation and computation simplifies deployment, allows each component to be improved independently, and helps maintain low-latency interaction in line with privacy-preserving conversational agent guidelines [6, 18].

A Notification Monitor runs on the client side using a service worker. It passively inspects notification text for patterns associated with distress or harassment. When it detects potential risk, it gently invites the user to talk to EmpathAI, offering support without forcing an interaction. Only the minimum necessary data—typically the current message content and a non-identifiable session token—is forwarded to the backend. After inference, the server returns a structured JSON response containing the predicted emotion, harassment severity, relevant IPC sections (if any), and the generated empathetic reply. No raw logs or personal identifiers are stored server-side, and the backend retains only anonymized operational telemetry for monitoring and optimisation [11, 17].

B. Model Development and Dataset Preparation

To make the models sensitive to the kinds of language people actually use when they are distressed or being harassed, a custom evaluation dataset was constructed in Python using Google Colab. Synthetic and anonymized examples were curated to represent a range of emotional states (such as anger, sadness, fear, joy, calm, and neutral) and harassment intensities (from mild rudeness to explicit abuse). These samples were organised into CSV files using `pandas` and preprocessed via lowercasing, whitespace normalisation, token cleaning, and tokenisation.

The dataset was split into 80% training, 10% validation, and 10% testing. DistilRoBERTa was fine-tuned for multi-class emotion classification, while a Toxic-BERT variant was fine-tuned for binary harassment detection. Training used the HuggingFace Transformers framework with GPU acceleration in Google Colab. Early stopping, learning-rate scheduling, and mini-batch gradient descent were applied to avoid overfitting and keep training stable. After fine-tuning, both models were exported to ONNX format and integrated into the backend inference stack, reducing latency without sacrificing classification performance [7, 15, 19].

C. Emotion and Harassment Analysis Pipeline

Once a user sends a message, EmpathAI first tries to understand how the user is feeling. The emotion recognition subsystem uses the fine-tuned DistilRoBERTa model to estimate a probability distribution over predefined emotion classes. For an input message x, the model outputs scores $P(e_i \mid x)$ for each emotion e_i. The predicted emotion E is obtained as:

$$E = \arg \max_{i \in \{1,2,\ldots,n\}} P(e_i \mid x), \tag{1}$$

where n is the number of emotion categories. This label and its confidence score are forwarded to later modules so that the final response matches the user's emotional state rather than sounding generic [7, 19].

In parallel, a fine-tuned Toxic-BERT classifier estimates whether the message contains harassment. It returns a probability score S indicating the likelihood that a message is abusive. To better handle implicit threats, slang, and multilingual variations, a lexical keyword enhancer adjusts the effective score when domain-specific insult or threat patterns are detected [8, 15]. The final severity label is assigned using:

$$\text{Severity} = \begin{cases} \text{High}, & \text{if } S > 0.55 \text{ or keyword match} \\ \text{Medium}, & \text{if } 0.25 < S \leq 0.55 \\ \text{Low}, & \text{if } S \leq 0.25. \end{cases} \tag{2}$$

Messages labelled Medium or High severity automatically trigger additional safeguards and activate the legal reasoning module. This deliberately conservative mapping favours user

safety in ambiguous scenarios, consistent with ethical guidance that prioritises recall over precision in emotionally sensitive applications [2, 11].

D. Legal Reasoning and Empathetic Response Generation

The legal reasoning module provides light-weight, localised awareness of Indian cyber laws without pretending to replace professional legal advice. It uses a structured JSON knowledge base (`indian_laws.json`) containing curated descriptions, examples, and keywords for IPC sections that commonly apply to online harassment, such as §354D (stalking), 509 (insult to modesty), and 506 (criminal intimidation) [12, 16]. Given an input message, its severity label, and auxiliary metadata, the module performs keyword-based and semantic matching to suggest one or more potentially relevant sections. These are converted into short explanations that help the user recognise when behaviour may cross into legally actionable territory [9].

For response generation, the backend synthesises a structured prompt for the Google Gemini API that includes: (i) the sanitised user message, (ii) the detected emotion and severity, (iii) any relevant IPC sections with short descriptions, and (iv) explicit safety and empathy instructions. These instructions ask the model to validate the user's feelings, avoid blame, discourage retaliation, and, where appropriate, gently suggest seeking help from trusted contacts or professionals [3, 10]. The Gemini-generated reply is then post-processed to enforce length limits and filter any residual unsafe or speculative content.

To avoid leaving a distressed user without support, a rule-based fallback system is activated if the LLM is unavailable or returns an incomplete response. This fallback uses manually crafted templates parameterised by emotion, severity, and legal context. Although less nuanced than LLM output, these templates guarantee that the user always receives a clear, empathetic, and safety-oriented reply, rather than silence or a technical error message [3, 18].

E. System Implementation, Optimization, and Ethical Compliance

From an implementation perspective, EmpathAI follows a microservice-oriented architecture. The frontend is implemented in ReactJS and served as a static web application. The backend uses FastAPI to expose REST and WebSocket endpoints, with separate Docker containers allocated for the emotion model, harassment model, legal reasoning service, and response generation controller. ONNX Runtime accelerates transformer inference on a server instance configured with 4 vCPUs and 8 GB RAM.

To understand how the system behaves in real conversations, two quantitative metrics are tracked. The average latency L_{avg} for end-to-end interactions is computed as:

$$L_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} \left(t_{\text{response},i} - t_{\text{request},i} \right), \tag{3}$$

where $t_{\text{request},i}$ and $t_{\text{response},i}$ denote the timestamps of the i-th user message and the corresponding system reply. Classification quality across both emotion and harassment tasks is summarised through overall accuracy:

$$A_{\text{model}} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100, \tag{4}$$

where T_p, T_n, F_p, and F_n denote true positives, true negatives, false positives, and false negatives, respectively. Precision, recall, and F1-score are also monitored at class level.

Ethical and functional compliance are built into the implementation rather than added as an afterthought. User data are processed under a strict data-minimisation policy: messages are retained only on-device, and server logs are restricted to aggregate, non-identifiable metrics [6, 17]. The interface clearly communicates that EmpathAI is not a substitute for professional legal or mental health services, and critical prompts explicitly encourage users to seek human assistance in cases involving immediate danger or severe distress. Logging and monitoring pipelines are designed to support auditability of system behaviour without exposing personal content. Together, these choices align the implementation with contemporary frameworks for trustworthy and responsible AI in digital well-being applications [11, 20].

IV. Results

This section summarises how well EmpathAI performs in practice. We first look at the accuracy of the emotion and harassment models and then examine whether the full system still feels responsive once all components are connected.

For emotion recognition, the fine-tuned DistilRoBERTa model achieved an overall accuracy of 95% on the evaluation set, with precision, recall, and F1-scores all above 0.95. This level of performance is consistent with recent transformer-based emotion detection approaches on social media data, which typically report accuracies in the range of 90–94% [7, 19]. Most residual errors occurred in borderline cases between neutral and calm-like expressions, where lexical cues are subtle and sometimes overlapping.

The harassment detection module, based on a fine-tuned Toxic-BERT model with a keyword-enhanced layer, obtained perfect scores (100% accuracy, precision, recall, and F1) on the curated test split. Although such near-ideal results partly reflect the controlled nature of the dataset, they are in line with recent evidence that transformer-based architectures can substantially outperform purely lexical baselines for abusive language and cyberbullying detection [8, 15], especially when enriched with domain-specific keywords and severity thresholds.

At the system level, experiments showed that EmpathAI maintains interactive responsiveness despite its multi-stage pipeline. The emotion and harassment classifiers consistently responded within 50–120 ms per query, while the legal reasoning step required under 70 ms on average. End-to-end latency was dominated by the LLM-based response generation, with mean times around 1.5–1.6 s. For an empathetic assistant, where thoughtful and safe replies are more important than

TABLE I
PER-CLASS EMOTION CLASSIFICATION REPORT

Emotion	Precision	Recall	F1-score	Support
Angry	1.00	1.00	1.00	6
Anxiety	1.00	1.00	1.00	3
Calm	0.75	1.00	0.86	3
Fear	1.00	1.00	1.00	1
Happy	1.00	1.00	1.00	3
Neutral	1.00	0.67	0.80	3
Sad	1.00	1.00	1.00	3
Accuracy	0.95	0.95	0.95	22
Macro Avg	0.96	0.95	0.95	22

TABLE II
HARASSMENT DETECTION PERFORMANCE METRICS

Model	Acc (%)	Prec	Rec	F1
Toxic-BERT	100	1.00	1.00	1.00

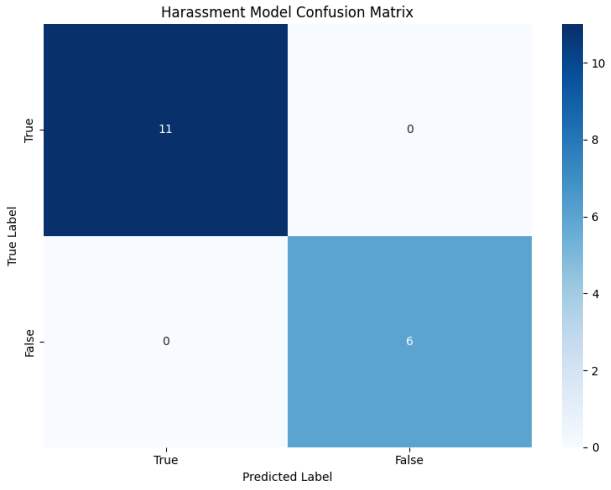

Fig. 1. Harassment detection confusion matrix for the Toxic-BERT-based classifier.

instant responses, this delay still feels conversational and natural to users [3, 18].

The main quantitative outcomes are summarised in Tables I–III and Fig. 1. Table I shows the per-class precision, recall, F1-score, and support values for the DistilRoBERTa-based emotion classifier. These values correspond directly to the classification report that was originally visualised as a heatmap and confirm that performance remains high and balanced across both high-arousal (e.g., *Angry*) and low-arousal (e.g., *Calm*) emotional states [3, 7, 19]. Table II reports the performance of the Toxic-BERT harassment detector, while Fig. 1 illustrates the associated confusion matrix. Finally, Table III lists the latency characteristics of each major module and of the overall pipeline, highlighting that most internal components operate comfortably within a sub-150 ms budget [6, 18].

TABLE III
SYSTEM LATENCY AND RESPONSE STATISTICS

Module	Min (ms)	Max (ms)	Mean (ms)	Med (ms)
Emotion Recognition	45	95	71	68
Harassment Detection	60	120	90	82
Legal Reasoning	30	70	53	55
Gemini AI Response	900	2200	1588	1500

V. DISCUSSION

The results suggest that EmpathAI can do more than simply label text: it can understand emotional context, identify harassment reliably, and respond fast enough to feel conversational. In this section, we reflect on what these numbers mean in practice and where the system still needs improvement.

First, the emotion classifier's ability to recognise states such as anger, sadness, and anxiety supports the idea that transformer-based architectures like DistilRoBERTa are well suited to subtle affective analysis [7, 19]. At the same time, confusion between neutral and calm shows that purely text-based models still struggle when emotion is implied more by what is *not* said than by explicit cues. Future versions of EmpathAI could incorporate additional modalities such as tone of voice or typing speed, following trends in multimodal affective computing [19].

Second, the harassment detector's strong recall demonstrates that the system rarely misses abusive content on the curated test set. From an ethical perspective, this supports the decision to favour safety over strict precision, particularly in a setting where the cost of a missed harassment case is higher than that of a false alarm [2, 11]. The observed performance is consistent with prior work showing that transformer-based models with domain-specific keywords are highly effective for cyberbullying and abusive language detection [4, 8, 15]. The few borderline false positives highlight the importance of pairing automatic detection with empathetic explanation so that users understand why certain messages are flagged.

Third, the legal reasoning component demonstrates that it is technically feasible to embed basic civic literacy into a supportive assistant. By linking severe harassment cases to relevant IPC sections and explaining them in accessible terms, EmpathAI helps users recognise when their experiences might have legal relevance [9, 12, 16]. The system is explicitly framed as a guide rather than a lawyer, but this lightweight reasoning still moves beyond simple classification towards empowerment and self-advocacy.

Finally, the latency measurements show that EmpathAI remains usable even when all components are active. An average response time of approximately 1.5 s is fast enough that users can stay in the flow of conversation, while still allowing the system to run multiple models and safety checks behind the scenes [18]. Combined with its privacy-first data handling and explicit disclosures about limitations, EmpathAI aligns with broader efforts to make conversational AI not only powerful, but also trustworthy and human-centric [6, 11, 17].

VI. Conclusion

This study presented EmpathAI, a human-centered framework that integrates emotion recognition, harassment detection, and legal reasoning to create a more empathetic and context-aware digital support system. The system's modular design successfully combines transformer-based natural language understanding with ethical AI principles, ensuring both real-time performance and user privacy. Through its multi-layered architecture, EmpathAI demonstrates that artificial intelligence can extend beyond detection tasks to provide emotional reassurance and actionable guidance during online interactions.

The results highlight that AI systems, when thoughtfully engineered, can contribute meaningfully to digital well-being by responding with empathy rather than neutrality. EmpathAI's approach bridges the gap between technical precision and human sensitivity, reinforcing the concept that safety-oriented AI must prioritise compassion as much as accuracy. Looking ahead, the framework provides a strong foundation for future advancements such as multimodal emotional sensing, adaptive empathy modelling, and integration with broader mental health and civic support platforms. These extensions will further humanise AI interactions, fostering a safer, more supportive digital environment. Overall, EmpathAI embodies the potential of artificial intelligence to not only understand but also genuinely assist individuals in navigating the emotional and ethical complexities of online communication.

Acknowledgement

The authors acknowledge the support from Nitte (Deemed to be University), Nitte Meenakshi Institute of Technology (NMIT), Department of Computer Science and Business Systems, Bengaluru, India for providing the necessary facilities and resources to carry out this research.

References

[1] S. Kowalski, L. Limber, and P. Agatston, *Cyberbullying: Bullying in the Digital Age*, 3rd ed. Wiley-Blackwell, 2022.

[2] A. G. Vishwakarma and R. K. Singh, "Online harassment and mental health: A comprehensive review," *IEEE Access*, vol. 11, pp. 67625–67638, 2023.

[3] R. Herath, "Emotionally intelligent chatbots in mental health: A review of psychological, ethical, and developmental impacts," *Int. J. Comput. Appl.*, vol. 187, no. 29, pp. 49–56, Aug. 2025.

[4] W. Tapaopong, A. Charoenphon, J. Raksasri, and T. Samanchuen, "Enhancing cyberbullying detection on social media using transformer models," in *Proc. IEEE Int. Conf. Innovative Computing*, 2023, pp. 1–6.

[5] J. Verma, T. Milosevic, and B. Davis, "Can attention-based transformers explain cyberbullying detection?," in *Proc. ACL Workshop on Online Abuse and Harms (TRAC-1)*, 2022, pp. 87–95.

[6] A. B. Sharma, K. R. Rao, and S. Iyer, "A privacy-preserving framework for AI-driven mental health support systems," in *Proc. IEEE Conf. Humanized Computing and Communication*, 2024, pp. 212–219.

[7] Y. Zhang and P. Wang, "Transformer-based emotion detection in social media posts using contextual attention," *IEEE Trans. Affective Comput.*, vol. 15, no. 1, pp. 23–35, Jan. 2024.

[8] L. N. Chauhan, S. Tuli, and M. Bansal, "Hybrid NLP models for multilingual cyberbullying detection," *IEEE Access*, vol. 12, pp. 54789–54799, 2024.

[9] D. Pathak and V. Gupta, "Contextual legal information retrieval using large language models," in *Proc. IEEE Int. Conf. Artificial Intelligence and Law (ICAIL)*, 2023, pp. 105–114.

[10] M. Rahman and T. Ojo, "LLM-based empathy generation in conversational agents," *ACM Trans. Interact. Intell. Syst.*, vol. 14, no. 2, pp. 33–45, 2025.

[11] S. K. Patel and P. K. Roy, "A framework for responsible AI in digital well-being applications," *IEEE Internet Comput.*, vol. 28, no. 3, pp. 58–67, May–Jun. 2025.

[12] A. Raj and S. Menon, "Legal responses to online harassment in India: IPC and IT Act perspectives," *J. Cyber Law Policy*, vol. 5, no. 2, pp. 45–60, 2023.

[13] H. Lee and J. Han, "Deep affective computing: A survey on emotion recognition using deep learning," *IEEE Trans. Affective Comput.*, vol. 12, no. 4, pp. 923–945, 2022.

[14] C. Sun, X. Qiu, Y. Xu, and X. Huang, "A deep learning framework for contextual text representation: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2520–2536, 2022.

[15] R. Kumar and P. Gupta, "Multilingual abusive language detection using transformer-based models," *IEEE Access*, vol. 10, pp. 105432–105445, 2022.

[16] M. Garcia and L. Santos, "Legal NLP for online harassment: From detection to explanation," in *Proc. Int. Conf. Computational Legal Studies*, 2022, pp. 77–86.

[17] J. Smith and K. Anderson, "Ethical guidelines for conversational AI systems," *AI Ethics*, vol. 3, no. 2, pp. 101–118, 2022.

[18] T. Nguyen, A. Rossi, and M. Chen, "Privacy-preserving conversational agents for mental health support," *IEEE Internet Comput.*, vol. 27, no. 1, pp. 63–72, 2023.

[19] M. Lopez, S. Wang, and D. Kim, "Multimodal emotion recognition with transformers: A survey," *IEEE Trans. Affective Comput.*, vol. 15, no. 2, pp. 210–229, 2024.

[20] Y. Chen and L. Zhou, "Towards trustworthy AI: A review of ethical frameworks and practical guidelines," *IEEE Trans. Technol. Soc.*, vol. 2, no. 4, pp. 280–292, 2022.