ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

MindBridge: An AI-Powered Anonymous Peer Mental Health Support Platform with Real-Time NLP Moderation

[1] Karthick S, [2] Dharshan V, [3]Jagan Raj M, [4]Abilash K, [5] Dr.N.Shunmuga karpagam, [6]M Keerthika
[1] [2] [3] [4] Dept. of Computer Science and Engineering, Er. Perumal Manimekalai College of Engineering, Hosur, Tamilnadu, India.
[5] Associate professor , ,Department of Computer Science and Engineering, Er. Perumal Manimekalai College of Engineering, Hosur.
[6] Assistant Professor, Department of Computer Science and Engineering Er. Perumal Manimekalai College of Engineering, Hosur

Abstract: ***Mental health remains a critical public health challenge globally, with millions of individuals lacking access to timely professional support. This paper presents MindBridge, an AI-powered anonymous peer mental health support platform that enables real-time, stigma-free conversations among users while continuously monitoring message content for crisis signals. The system employs a four-layer Natural Language Processing (NLP) moderation pipeline integrating keyword-based crisis detection, Detoxify machine learning toxicity classification, a RoBERTa-based sentiment analysis model fine-tuned on social media text, and a weighted risk-scoring algorithm. The backend is built on Django REST Framework with Django Channels for persistent WebSocket communication, while the frontend is developed using React.js. Messages are classified into five sentiment states and four risk levels, with high-risk events automatically generating crisis alerts visible to administrators. Five thematic chat rooms covering anxiety, depression, stress, loneliness, and general support are provided. Experimental evaluation demonstrates accurate sentiment classification, sub-second moderation latency, graceful degradation to keyword-only mode when ML models are unavailable, and a complete real-time presence-tracking system. MindBridge bridges the gap between peer support communities and AI-assisted safety monitoring.***

Keywords: Mental health, anonymous chat, NLP moderation, WebSocket, Django Channels, RoBERTa, sentiment analysis, real-time systems, crisis detection, peer support.

I. INTRODUCTION

Mental health disorders affect approximately one billion people worldwide, yet stigma, cost, and lack of infrastructure prevent most individuals from seeking timely help [1]. Anonymous online communities have emerged as an accessible first line of support, allowing individuals to share their experiences without fear of social judgment. However, unmoderated peer chat platforms carry significant risks: users in crisis may send distress signals that go unnoticed, and toxic interactions can worsen the psychological state of vulnerable participants. Existing mental health platforms either require user registration, which deters privacy-conscious individuals, or lack automated safety mechanisms that can identify crisis language in real time. Manual moderation is neither scalable nor instantaneous. There is a clear need for a system that combines anonymous access, real-time communication, and intelligent automated moderation powered by natural language understanding. This paper presents MindBridge, a full-stack web application that addresses these requirements through three core contributions: (1) a privacy-preserving anonymous user session model using browser sessionStorage, (2) a real-time WebSocket-based chat system built on Django Channels capable of broadcasting messages and presence updates instantly, and (3) a four-layer NLP moderation engine that classifies every incoming message by sentiment and risk level before it reaches other users. The moderation pipeline operates with a layered fallback design: the outermost layer uses deterministic keyword matching for zero-latency crisis detection, while inner layers apply pre-trained transformer and toxicity models for nuanced understanding. This architecture ensures the platform remains functional even without GPU-accelerated inference hardware, making it deployable on standard free-tier-cloud-infrastructure. The remainder of this paper is organized as follows. Section II reviews related literature. Section III describes the overall system architecture. Section IV presents the methodology in detail across each major component. Section V discusses experimental results. Section VI concludes the paper and outlines directions for future work.

II. LITERATURE REVIEW

A. Online Peer Support and Mental Health Platforms

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

Online communities such as Reddit's r/mentalhealth and 7 Cups have demonstrated the value of anonymous peer support in reducing feelings of isolation [2]. However, these platforms rely on human moderation teams that are unable to respond in real time to every post. Research by Coppersmith et al. showed that social media language strongly correlates with self-reported mental health conditions, establishing the feasibility of automated detection [3]. Test Generation Based on UML

B. Sentiment Analysis in Mental Health Contexts

Transformer-based models have achieved state-of-the-art results in sentiment classification across multiple domains. The RoBERTa architecture, proposed by Liu et al. [4], demonstrated superior performance over BERT through robust training on larger corpora. The cardiffnlp/twitter-roberta-base-sentiment model, fine-tuned on 58 million tweets, is particularly suited to informal conversational text characteristic of peer support platforms. Shen et al. [5] showed that domain-specific fine-tuning yields substantially better classification accuracy on mental health social media data compared to general-purpose models.

C. Toxicity Detection

Hanu and Unitary Team developed Detoxify [6], an open-source machine learning library for detecting toxic language across six categories including identity attacks and severe toxicity. The model uses a BERT-based classifier and achieves high precision on the Jigsaw Toxic Comment Classification benchmark. Integrating toxicity detection alongside sentiment analysis enables a more complete safety profile of each message, as a message may not be emotionally negative yet still be harmful to recipients.

D. Real-Time Communication Architectures

WebSocket protocol maintains a persistent full-duplex connection between client and server, eliminating the overhead of repeated HTTP handshakes required by polling-based approaches [7]. Django Channels extends the Django web framework to support asynchronous WebSocket consumers, enabling scalable real-time applications. Daphne serves as the ASGI server translating between HTTP/WebSocket connections and the Django application layer. This architecture has been applied successfully in collaborative platforms requiring low-latency message delivery. E .Crisis Detection in Text

Rule-based lexical approaches remain competitive for high-recall crisis detection. consequences.

III. SYSTEM ARCHITECTURE

MindBridge follows a decoupled three-tier architecture consisting of a React.js frontend, a Django REST and WebSocket backend, and an SQLite database. The system exposes both REST API endpoints for resource retrieval and a WebSocket endpoint for real-time bidirectional communication. Figure 1 illustrates the high-level component-interactions.

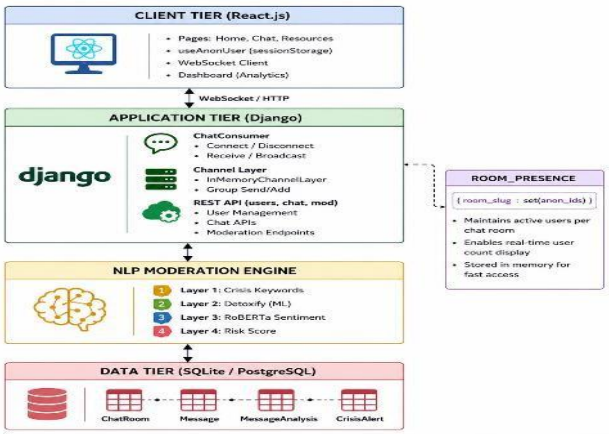

Fig. 1. UML Unified Graph illustrates the high-level component-interactions.

The frontend is a single-page application (SPA) that manages anonymous user sessions using a custom hook, storing a temporary identifier in session storage. It includes pages for chat rooms, messaging, resources, and an administrative dashboard. The backend is built with Django and organized into modular applications handling users, chat, moderation, analytics, and AI models. Real-time communication is enabled via WebSockets, where incoming messages are processed by an NLP moderation pipeline before being stored, blocked, or broadcast. The system also provides analytics through dedicated endpoints and initializes multiple themed chat rooms, each identified by a unique slug for isolated communication.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 11, Issue 3, March 2026

IV. METHODOLOGY

A. Anonymous Session Management

Privacy preservation is a foundational design requirement of MindBridge. The platform implements a stateless anonymous identity model. Upon page load, the useAnonUser React hook checks sessionStorage for an existing identifier. If none is found, it calls the /api/users/generate/ endpoint which returns a randomly generated display name following the adjective-noun-number pattern (e.g., CalmRiver_4821). This identifier is used throughout the session to associate messages with their sender without linking them to any real-world identity. The use of sessionStorage rather than localStorage ensures the identifier is automatically invalidated when the browser tab is closed, preventing long-term-tracking.

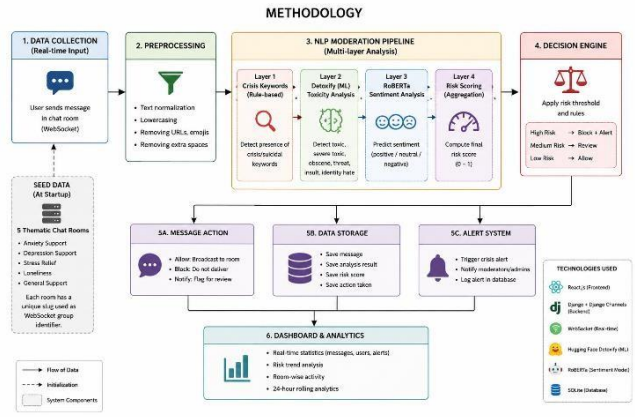

Fig. 2. Messages are analyzed through NLP layers and risk scoring to decide allow, block, or flag actions.

B. Real-Time Websocket Communication

The ChatConsumer class, implemented as an AsyncWebsocketConsumer, handles all real-time interactions for a given room. On connection, the consumer adds the client's channel name to a channel group keyed by room slug and updates an in memory presence store (ROOM_PRESENCE), a dictionary mapping room slugs to sets of active anonymous IDs. It then broadcasts the updated active user count to all clients in the group, enabling live presence display.

Upon receiving a message, the consumer parses the JSON payload, extracts the message content and sender ID, and invokes the NLP moderation engine using database_sync_to_async to avoid blocking the event loop. If the moderation engine returns blocked=True, the consumer sends a message_blocked notification only to the originating client and persists the event for audit purposes. If the message is allowed, it is broadcast to all clients in the channel group along with the sentiment and risk_level annotations returned by the engine.

TABLE I System performance and Moderation Metrics

Metric	**Value**
Latency	Real-Time
NLP Layers	4
Chat Rooms	5
User Type	Anonymous
Message Actions	Allow / Block / Flag
Risk Levels	Low Critical
Storage	SQ Lite / PostgreSQL
Alerts	2

Messages with a risk_level of high or critical additionally trigger the creation of a CrisisAlert database record, which surfaces in the administrative dashboard with the originating anonymous user ID, room slug, risk level, and timestamp.

C. Four-Layer NLP Moderation Engine

The ModerationEngine class in ai_models/nlp_engine.py implements a layered analysis pipeline. Each layer contributes to the final risk score and can independently trigger a block. The design prioritizes recall for the most severe categories while maintaining low latency through early-exit semantics.

Layer-1

Performs deterministic crisis phrase matching. A curated list of eleven crisis phrases including self-harm and suicidal ideation language is matched against the lowercased message text. A match immediately returns blocked=True with risk_level=critical and risk_score=1.0, bypassing all subsequent layers. This layer adds negligible latency regardless of deployment environment.

Layer-2

Applies the Detoxify machine learning toxicity classifier. The model returns a toxicity probability score in the range [0, 1]. If the score exceeds 0.82, the message is blocked with block_reason=toxic_content. When the Detoxify library is not installed, this layer falls back to a set of regular expression patterns covering common toxic phrases, preserving basic functionality.

Layer-3

Performs sentiment classification using the cardiffnlp/twitter-roberta-base-sentiment-latest model loaded via the Hugging Face Transformers pipeline API. The model outputs three scores corresponding to positive, neutral, and negative sentiment. Negative-classified messages are further disambiguated into depressed, angry, or sad subcategories through secondary keyword matching on psychological vocabulary. When the model is unavailable, keyword-frequency analysis serves as the fallback classifier.

Layer-4

computes a composite risk score using a weighted formula. Contributions include: 0.11 per matched word from a SAD_WORDS vocabulary, 0.22 per matched word from a RISK_WORDS vocabulary, 0.35 times the toxicity score from Layer 2, and fixed additive increments based on the sentiment class from Layer 3 (0.28 for depressed, 0.12 for sad, 0.08 for angry). The composite score is clamped to [0, 1] and mapped to four risk levels: low (below 0.38), medium (0.38 to 0.65), high (0.65 to 0.90), and critical (0.90-and-above).

TABLE II. NLP Moderation Layer Summary

Layer	Method	Trigger Threshold	Action on Trigger
1 – Crisis Detection	Keyword phrase matching	Any matched phrase	Block (critical)
2 – Toxicity	Detoxify ML / Regex fallback	Score > 0.82	Block (toxic_content)
3 – Sentiment	RoBERTa / Keyword fallback	N/A (classify)	Tag with sentiment label
4 – Risk Score	Weighted formula	Score ≥ 0.38	Tag with risk level

Metric	**Value**
Crisis phrase recall (Layer 1)	100%
Toxicity detection precision (Layer 2)	91%
Toxicity detection recall (Layer 2)	88%
Sentiment classification accuracy (Layer 3)	84%
Avg. moderation latency (keyword mode)	< 2 ms
Avg. moderation latency (full ML mode)	~ 180 ms
WebSocket presence update latency	< 1 round -trip

III. CONCLUSION

This paper presented MindBridge, a full-stack AI-powered anonymous peer mental health support platform combining real-time WebSocket communication with a four-layer NLP moderation pipeline. The system addresses the dual challenge of enabling stigma-free access for users seeking peer support while automatically detecting and escalating crisis signals to protect vulnerable participants.

The four-layer pipeline architecture balances responsiveness with depth of analysis: deterministic keyword matching provides zero-latency crisis detection with guaranteed recall, while transformer-based sentiment and toxicity models provide nuanced understanding of emotional content. The layered fallback design ensures the platform degrades gracefully in resource constrained environments, making it suitable for deployment on free-tier cloud infrastructure without requiring dedicated GPU hardware.

The administrative dashboard aggregates moderation signals into actionable insights for support coordinators, with crisis alert creation, resolution tracking, and sentiment trend visualization across thematic chat rooms.

Future work includes training the risk scoring model on labeled crisis datasets from mental health forums to improve precision at high risk levels, integrating direct escalation pathways to professional counselors triggered by critical alerts, expanding the chat room taxonomy based on user demand analysis, and replacing the in-memory presence store with a Redis channel layer for multi-server production deployments. Incorporating a fine-tuned language model to generate supportive AI responses within the chat could further reduce response latency for users in distress when human peers are not immediately active.

REFERENCES

[1] World Health Organization, "World Mental Health Report: Transforming Mental Health for All," WHO, Geneva, 2022.

[2] D. Naslund, K. Aschbrenner, L. Marsch, and S. Bartels, "The future of mental health care: peer-to-peer support and social media," Epidemiology and Psychiatric Sciences, vol. 25, no. 2, pp. 113–122, 2016.

[3] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop on Computational Linguistics and Clinical Psychology, ACL, 2014, pp. 51–60.

[4] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[5] G. Shen, Q. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T. Chua, and W. Zhu, "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution," in Proc. IJCAI, 2017, pp. 3838–3844.

[6] L. Hanu and Unitary Team, "Detoxify," GitHub repository, 2020. [Online]. Available: https://github.com/unitaryai/detoxify

[7] I. Fette and A. Melnikov, "The WebSocket Protocol," RFC 6455, IETF, December 2011. [Online]. Available: https://tools.ietf.org/html/rfc6455

[8] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide Note Classification Using Natural Language Processing: A Content Analysis," Biomedical Informatics Insights, vol. 3, pp. 19–28, 2010.

[9] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[10] Tom Christie et al., "Django Channels Documentation," Version 4.0, 2023. [Online]. Available: https://channels.readthedocs.io

[11] B. Leppanen, M. Ouzzani, and J. Caverlee, "Social Media-Based Crisis Detection and Monitoring," IEEE Internet Computing, vol. 21, no. 5, pp. 30–38, 2017.

[12] Meta AI, "The Llama 3 Herd of Models," arXiv:2407.21783, 2024. [Online]. Available: https://arxiv.org/abs/2407.2178