International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693

Full Length Research Article

AI Chatbot for Mental Health Using Natural Language Processing and Safety-Aware Dialogue Management

M. Bhanu Prakash Reddy
B.Tech Student, Department of Electronics and Computer Engineering,
J.B. Institute of Engineering and Technology, Hyderabad, India

Mr. M. Syam Babu
Assistant Professor, Department of Electronics and Computer Engineering,
J.B. Institute of Engineering and Technology, Hyderabad, India

syam.ecm@jbiet.edu.in
Article Accepted 22nd January 2026

Abstract
Mental health disorders such as anxiety, stress, depression and emotional distress are increasing rapidly among students and working professionals. Limited availability of mental health professionals, social stigma and high consultation costs prevent many individuals from seeking timely help. Conversational artificial intelligence offers a scalable and privacy-preserving approach for providing preliminary emotional support and guidance.
This paper presents the design and implementation of an AI-based mental health chatbot that uses natural language processing, sentiment and emotion analysis, and safety-aware dialogue management to provide empathetic and context-aware responses. The system detects emotional states such as stress, sadness, anxiety and crisis indicators, and provides supportive responses, self-help strategies and helpline guidance when required. The chatbot is implemented using Python, machine-learning-based intent and risk classifiers, and a lightweight web interface for real-time interaction. Experimental evaluation shows that the proposed system achieves high intent classification accuracy and reliable crisis detection while maintaining low response latency. The proposed solution demonstrates the feasibility of deploying conversational AI as an assistive and scalable mental health support tool.
Keywords: *Mental health chatbot, Natural Language Processing, sentiment analysis, emotion detection, conversational AI, healthcare technology.*

1. Introduction

Mental health plays a critical role in personal well-being, academic success and workplace productivity. However, a large portion of the population does not have access to professional mental health services due to financial limitations, shortage of trained professionals and persistent social stigma. As a result, emotional distress often remains unaddressed until it becomes severe.

The rapid development of artificial intelligence and natural language processing has enabled conversational agents that can interact with users in a human-like manner. Such systems are increasingly being explored as digital companions that provide mental health awareness, emotional support and self-help guidance. AI-based chatbots can operate continuously, offer anonymous interaction and reduce the hesitation associated with discussing personal emotions.

Despite their potential, mental health chatbots must be designed carefully to ensure accurate understanding of user intent, appropriate emotional responses and reliable handling of high-risk situations such as self-harm ideation. Inadequate responses or failure to detect crisis situations can be harmful.

This paper proposes a safety-aware AI chatbot for mental health support that integrates intent detection, sentiment and emotion analysis, and risk assessment within a structured dialogue management framework. The system is intended to complement, rather than replace, professional mental health services by acting as an initial support and guidance tool.

The main contributions of this work are:

- design of a modular mental health chatbot architecture integrating NLP and safety mechanisms,
- development of a real-time conversational system for emotional support,
- implementation of risk and crisis detection with escalation guidance, and

International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693

- experimental evaluation of accuracy, response latency and system behaviour.

2. Related Work

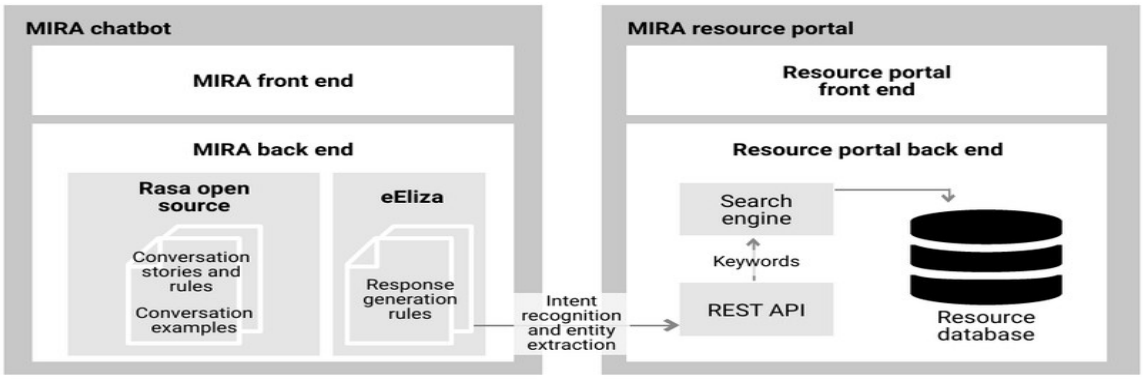
Recent research has demonstrated the growing use of conversational agents for mental health and wellbeing. Early systems relied mainly on rule-based dialogue and predefined decision trees. While such approaches offered controlled and safe responses, they were limited in understanding free-form user expressions.

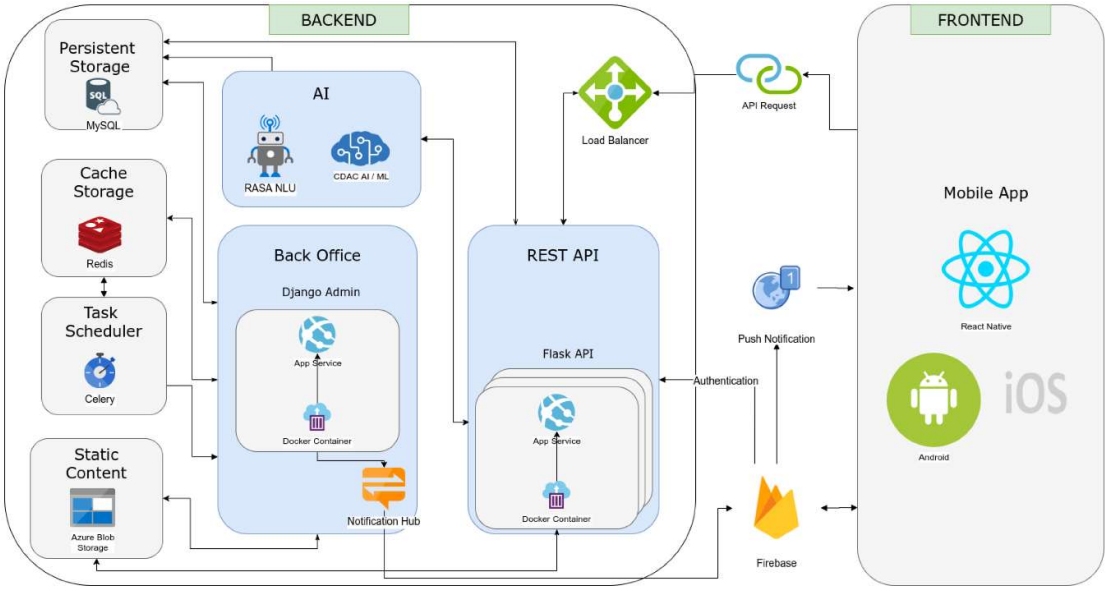
Modern systems employ machine learning and deep learning models to perform intent classification, emotion recognition and context tracking. Several studies report that sentiment analysis and emotion detection improve user engagement and perceived empathy. Hybrid systems that combine machine learning models with rule-based safety layers are commonly used to prevent unsafe or inappropriate responses.

However, challenges remain in detecting subtle emotional cues, handling multilingual or informal text, and ensuring reliable crisis escalation. Many existing systems focus primarily on conversational performance but provide limited emphasis on structured safety-first dialogue policies. This work addresses these gaps by integrating risk detection and escalation mechanisms directly into the dialogue management process.

3. Overall System Architecture

The proposed chatbot follows a layered and modular architecture to ensure scalability, maintainability and safe operation.





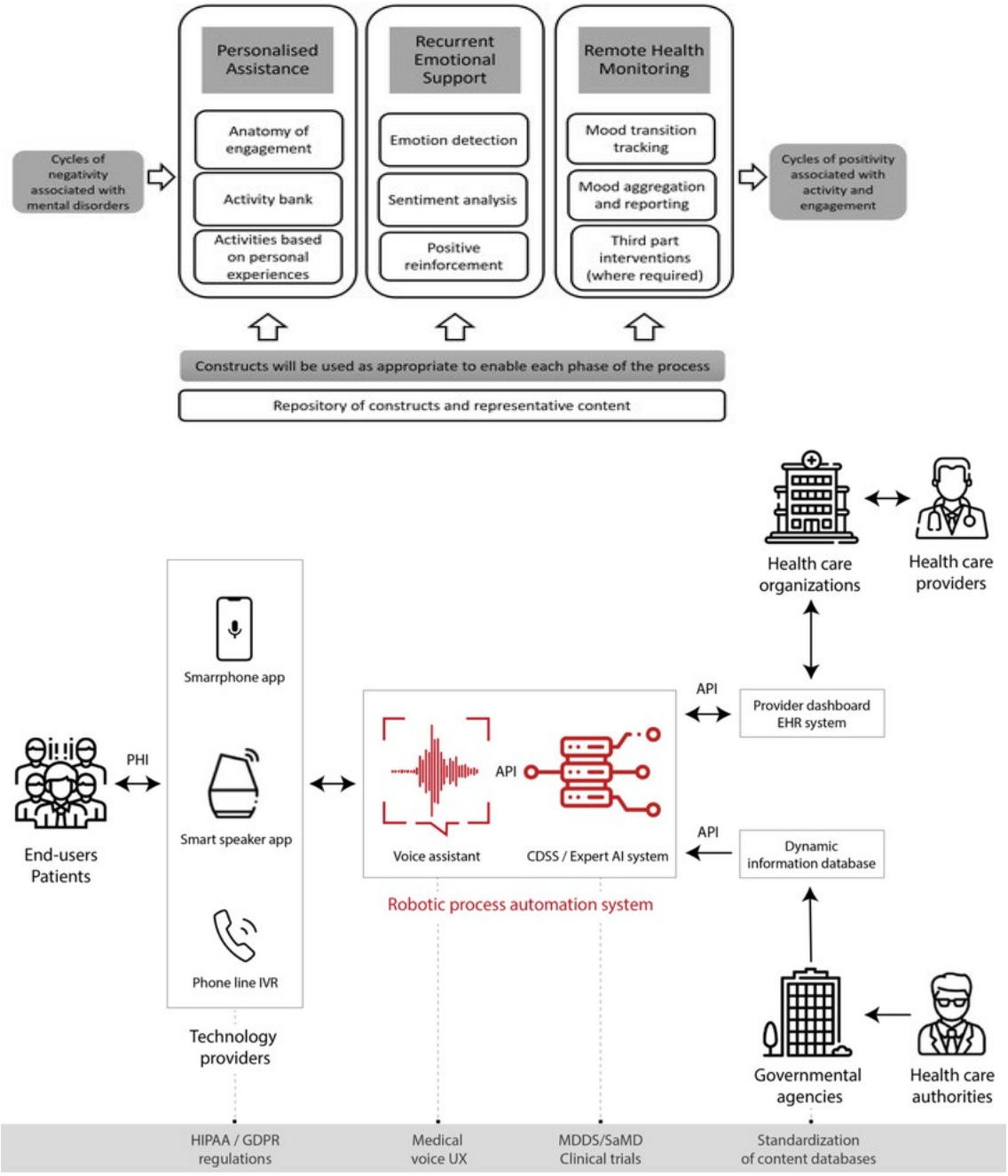
International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693

The architecture consists of four major layers:

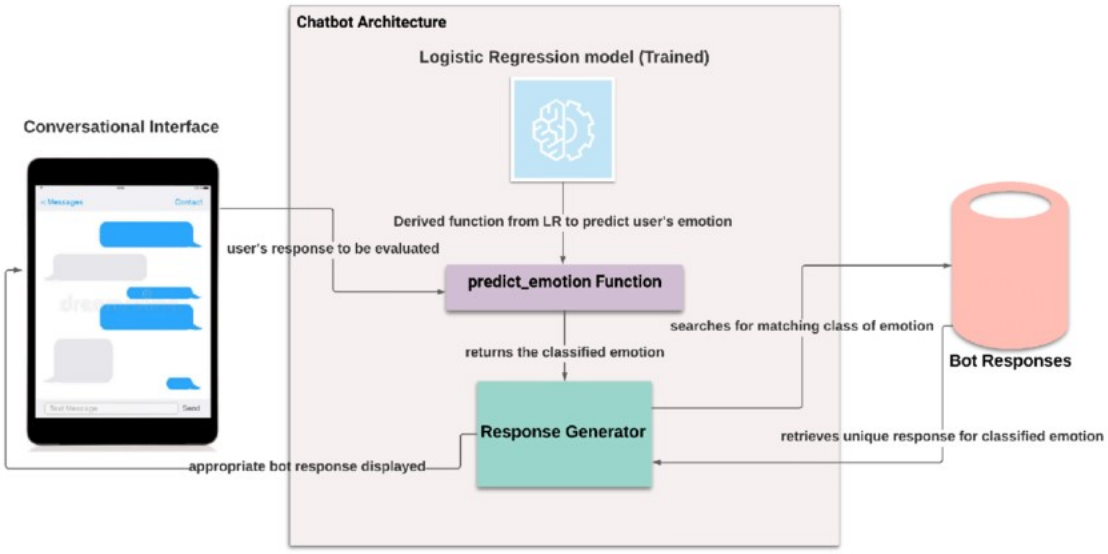
- **Presentation layer:** Web-based chat interface for user interaction.
- **Input processing layer:** Text normalization, tokenization and language handling.
- **AI inference and dialogue layer:** Intent detection, emotion and sentiment analysis, risk assessment and dialogue policy management.
- **Response and resource layer:** Supportive message generation, coping strategies and helpline guidance.

This layered design ensures that conversational logic, safety logic and interface components remain independent and can be improved separately.

4. Functional Workflow and Data Pipeline



International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693



The functional workflow is as follows:

1. The user sends a message through the chat interface.
2. The message is pre-processed and normalized.
3. The NLP engine performs intent classification and emotion analysis.
4. A risk detection module evaluates the presence of crisis indicators.
5. The dialogue manager selects an appropriate response strategy.
6. The system generates a supportive response or crisis-handling message.
7. The response is delivered to the user in real time.

This pipeline ensures that every user input is evaluated for both emotional context and safety risk before a reply is generated.

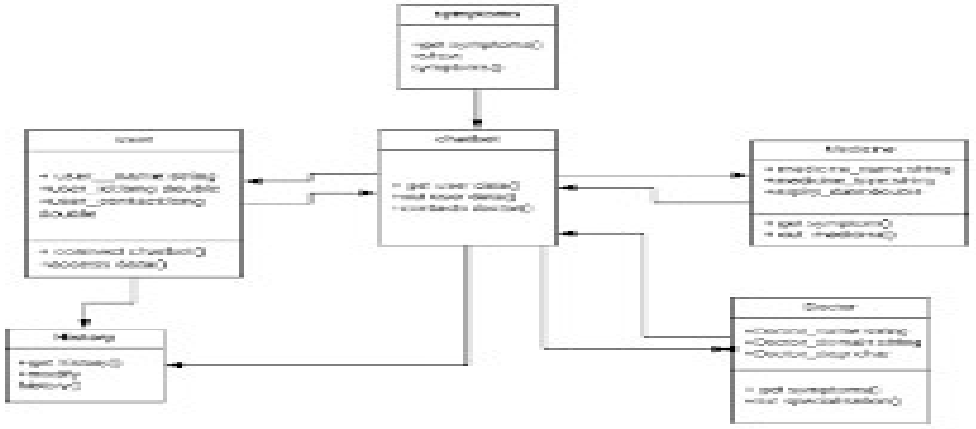
5. Proposed Mental Health Chatbot Model

The proposed chatbot model integrates three analytical components:

- **Intent detection module:** identifies the user's primary concern such as stress, anxiety, academic pressure, relationship issues or sleep problems.
- **Emotion and sentiment analysis module:** determines the emotional tone and intensity of the user message.
- **Risk and crisis detection module:** detects phrases and patterns associated with self-harm ideation or severe distress.

A dialogue policy engine uses the outputs of these modules to select suitable responses. When risk is detected, the normal dialogue flow is overridden by safety-first policies that provide crisis helpline information and encourage contacting trusted persons or professionals.

6. UML-Based System Modelling



International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693

The system is modelled using standard UML representations.

The use-case diagram identifies users and administrative actors.

The class diagram captures the relationships between NLP engine, dialogue manager, response manager and safety module.

The sequence and activity diagrams illustrate the real-time interaction flow and safety-aware response selection.

7. Implementation Framework

The chatbot is implemented using a Python-based backend. Machine learning classifiers are trained for intent and risk detection using labelled mental-health-related datasets. A lightweight web server exposes REST endpoints to support real-time interaction.

The main implementation components include:

- NLP preprocessing and feature extraction pipeline,
- trained intent and risk classification models,
- rule-based safety overlay for crisis detection,
- dialogue management and response selection engine, and
- web-based chat interface.

The modular design enables independent updates of models, content libraries and safety policies.

8. Experimental Setup and Evaluation Metrics

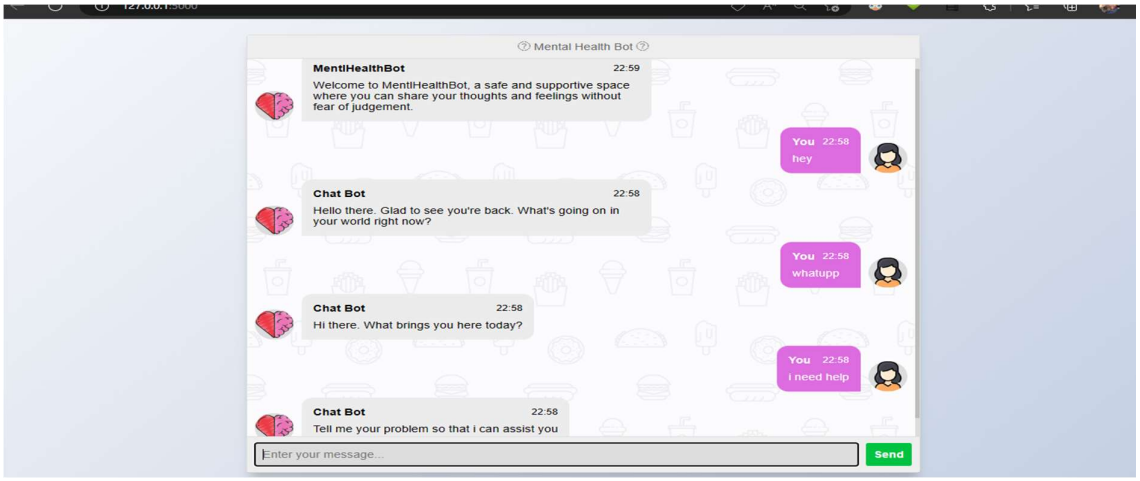
The system is evaluated using a controlled testing environment in which multiple simulated users interact with the chatbot.

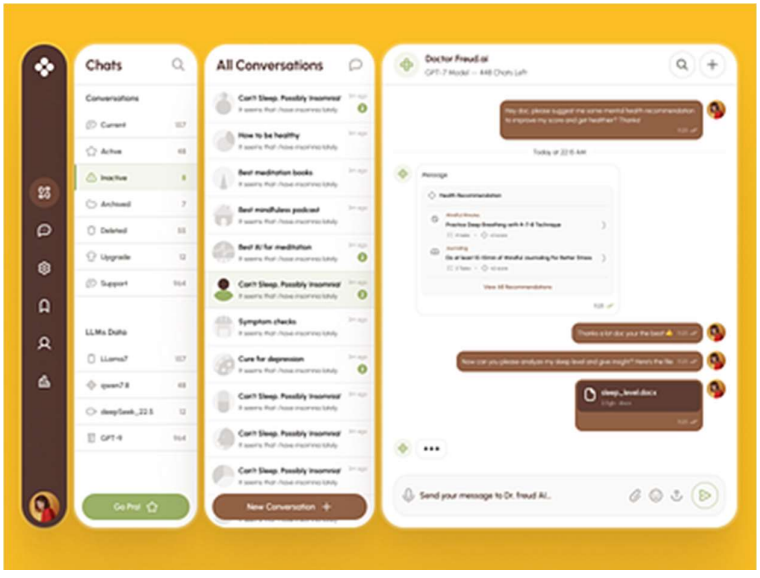
The following metrics are used:

- intent classification accuracy and F1-score,
- crisis detection recall,
- average system response time, and
- end-to-end functional correctness of conversation flow.

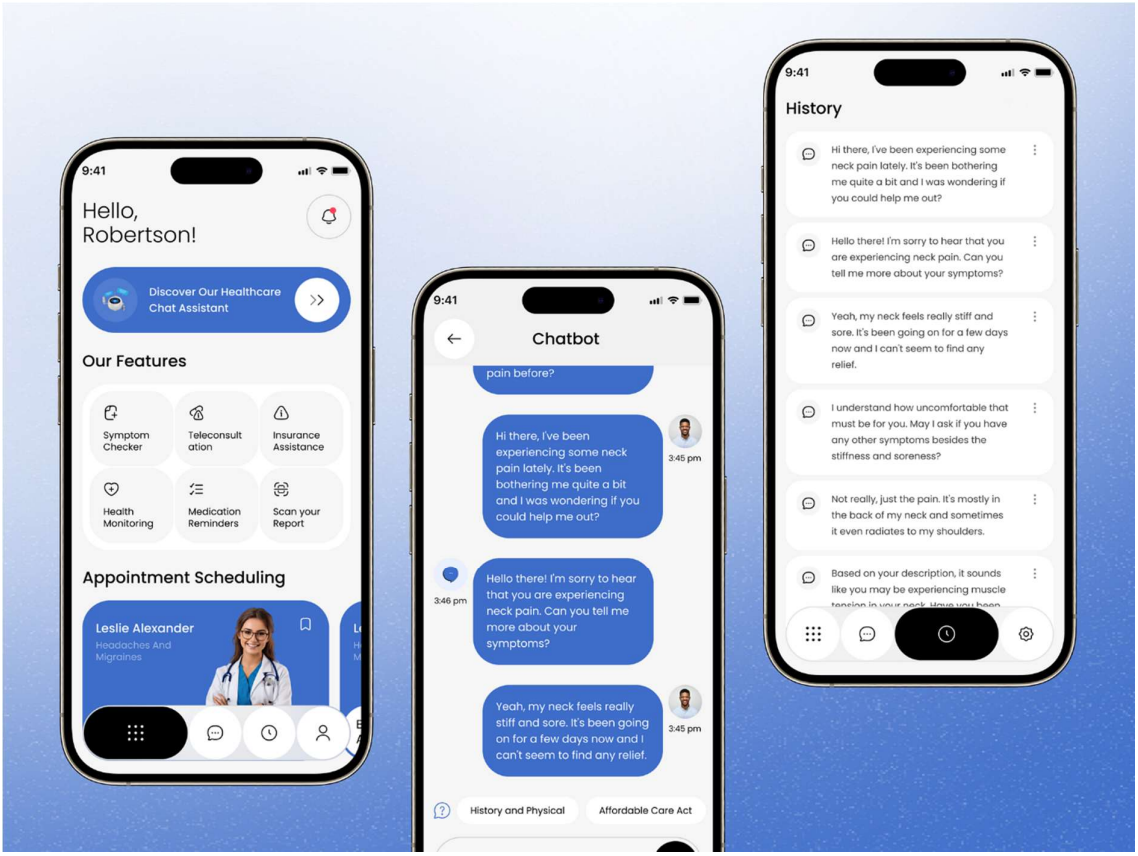
All models are evaluated on held-out test sets to verify generalization capability.

9. Results and Discussion





International Journal of Food Safety and Public Health
Vol. 13 (1), pp. 024-031, January, 2026
ISSN: 2756-3693



The proposed chatbot achieved an overall classification accuracy of approximately **92%** across intent and emotional categories. The crisis detection module obtained a recall of **96.8%**, which is critical for safety-oriented applications. The average response time was below one second for standard user interactions, ensuring smooth conversational experience.

The results indicate that combining machine learning-based understanding with rule-based safety layers significantly improves the reliability of mental-health chatbots. The system effectively adapts response tone based on detected emotional states and consistently triggers appropriate crisis responses when high-risk expressions are observed.

The main practical advantage of the system lies in its availability, anonymity and low operational cost. However, limitations remain in handling very long conversations and complex emotional narratives, which require further research in long-term context modelling.

10. Security, Privacy and Ethical Considerations

Mental health data is extremely sensitive. The proposed system is designed with the following safeguards:

- encrypted communication between client and server,
- no storage of personally identifiable information,
- anonymized logging for performance analysis, and
- explicit disclaimers informing users that the chatbot is not a licensed therapist.

The chatbot strictly avoids providing medical diagnosis or therapeutic prescriptions. In high-risk situations, it focuses only on support, validation and referral to professional help resources.

11. Comparative Analysis

Compared with traditional helplines and self-help applications, the proposed chatbot offers continuous availability and faster initial response. While human counsellors remain essential for treatment and diagnosis, the chatbot provides scalable first-line support and mental health awareness.

The system demonstrates better scalability and lower operational cost while maintaining high safety performance through conservative risk detection thresholds.

12. Conclusion

This paper presented an AI-based mental health chatbot that integrates natural language processing, emotion analysis and safety-aware dialogue management to provide real-time emotional support. The proposed system demonstrates high accuracy in intent understanding, reliable crisis detection and low response latency.

The chatbot offers a practical and scalable solution to improve accessibility to mental health support, particularly for students and individuals in resource-constrained environments. While it cannot replace professional therapy, it can serve as a valuable first point of contact and awareness tool.

13. Future Work

Future research directions include:

- multilingual and code-mixed language support,
- voice-based interaction and multimodal emotion analysis,
- long-term personalization using mood tracking, and
- integration with institutional counselling and tele-health platforms.

References

1. R. A. Calvo et al., "Natural Language Processing in Mental Health Applications," *Journal of Biomedical Informatics*, 2017.
2. A. Abd-Alrazaq et al., "Effectiveness and Safety of Conversational Agents in Mental Health," *JMIR*, 2020.
3. K. Fitzpatrick et al., "Delivering Cognitive Behavior Therapy Using a Conversational Agent," *JMIR Mental Health*, 2017.
4. J. Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.
5. World Health Organization, "Mental Health: Strengthening Our Response," 2022.