



Sentimentix: Mental Health Companion

Mohan Jha¹, Devanshu Munjal², Vanshika Jain³, Roshni⁴

Department of CSE-AI&DS,

Panipat Institute of Engineering & Technology, Samalkha, India

Mail ID: roshni.cse.ai-ds@piet.co.in

Abstract—Mental-health support systems increasingly use natural language processing (NLP) and conversational artificial intelligence to provide accessible, low-barrier assistance. This paper presents Sentimentix, a software-based mental health companion designed to analyze user messages, infer sentiment or emotional state, maintain lightweight conversational context, and return supportive responses. The proposed framework combines text preprocessing, sentiment/emotion classification, contextual state management, response generation, and a safety-oriented risk and intent layer. Unlike systems that treat every message identically, Sentimentix adapts its response strategy according to the detected emotional polarity and conversational context. The design emphasizes privacy-aware local processing where feasible, transparent response rules, and escalation guidance for high-risk expressions. A prototype-oriented evaluation is described using functional scenarios rather than claiming clinical efficacy. The paper discusses the strengths and limitations of sentiment-aware conversational support and identifies future directions including multilingual modeling, transformer-based emotion recognition, longitudinal mood analytics, human-in-the-loop escalation, and clinically validated evaluation. Sentimentix is intended as a supportive technology and not as a replacement for qualified mental-health professionals.

Keywords: mental health, conversational AI, sentiment analysis, natural language processing, emotion detection, chatbot, digital mental health, safety-aware AI.

INTRODUCTION

Mental health support is an area in which accessibility, continuity, privacy, and timely communication are important. Conversational agents have been investigated as a way of providing scalable, text-based interaction and structured support. Systematic reviews report promising short-term effects for some conversational-agent interventions, while also emphasizing the need for rigorous evaluation and appropriate safety mechanisms [1], [2]. Recent reviews further show that mental-health conversational-agent research spans rule-based systems, machine-learning models, and large language models, with substantial variation in evaluation practices [3], [4].

Sentiment analysis provides a useful computational layer for such systems because the same literal request may require different responses depending on the user's emotional state. A message expressing frustration, loneliness, anxiety, or positive affect can therefore be routed to different response strategies. Sentiment-aware mental-health chatbots have already been explored using NLP preprocessing and recurrent neural networks, demonstrating the feasibility of integrating emotional classification with conversational response generation [5].



The objective of Sentimentix is to develop an accessible prototype that uses the user's textual input as the primary interaction modality and produces context-sensitive supportive responses. The system is not designed to diagnose depression, anxiety, or any other clinical disorder. Instead, it seeks to recognize conversational signals, encourage reflection and coping-oriented actions, and provide safety-oriented guidance when messages contain potentially serious risk indicators.

The main contributions of this work are: 1) a modular sentiment-aware conversational architecture; 2) integration of sentiment/emotion classification with response selection; 3) a lightweight risk and intent layer for safer routing; 4) a privacy-conscious interaction design; and 5) a prototype evaluation framework that distinguishes technical performance from clinical effectiveness.

LITERATURE REVIEW

Conversational agents for mental health have evolved from scripted dialogue systems to data-driven and generative architectures. Gaffney et al. reviewed conversational-agent interventions and reported encouraging reductions in psychological distress across several studies, while noting heterogeneity in therapeutic approaches and study quality [6]. A later meta-analysis of randomized controlled trials found statistically significant short-term improvements across several outcomes, including depressive symptoms, generalized anxiety, distress, and stress, although effect sizes and study conditions varied [1].

Li et al. synthesized experimental evidence on AI-based conversational agents and found that user experience is influenced by communication quality, engagement, personalization, and the human–AI therapeutic relationship [2]. Their review also identified sentiment or emotion analysis, mood tracking, empathic responses, and personalization as recurring design features. These findings motivate the integration of an emotional-state component within Sentimentix.

Cho et al. reviewed mental-health conversational-agent research across computer science and medical literature and highlighted a disciplinary gap: technical studies often emphasize model and language metrics, whereas medical studies more frequently evaluate health outcomes [3]. This distinction is important for Sentimentix because a high sentiment-classification score alone cannot establish that a user is clinically better. The proposed evaluation therefore separates system-level behavior from claims about mental-health outcomes.

Recent work has also focused on NLP-enabled conversational agents based on cognitive behavioral therapy principles. A 2026 systematic review and meta-analysis found small-to-moderate effects for depressive symptoms and small effects for negative affect, while effects for several other outcomes were not significant after publication-bias adjustment [7]. These results support cautious optimism rather than unrestricted claims about AI counseling.

Sentiment analysis has been explicitly integrated into mental-health support chatbots. One recent IEEE conference study developed a hybrid Rasa chatbot with an LSTM sentiment component and reported 76% training/validation accuracy for its sentiment model [5]. The study illustrates a practical architecture in which preprocessing, sentiment inference, response generation, and support suggestions are connected. Sentimentix follows the same broad principle but places greater emphasis on contextual state, safety routing, and non-clinical positioning.

Safety remains a central research challenge. Reviews of AI mental-health systems have found inconsistent reporting of safeguards, external validation, and governance practices [4], [8]. Consequently, Sentimentix includes a risk/intent stage before response generation. The stage is intended to detect potentially high-



risk language and route the interaction toward supportive safety guidance rather than attempting automated diagnosis or treatment.

TABLE I.COMPARISON OF RELEVANT APPROACHES

Approach	Core Technique	Strength	Limitation
Rule-based chatbot	Predefined intents/rules	Predictable responses	Limited flexibility
Sentiment chatbot [5]	NLP + LSTM	Emotion-aware response	Model/data dependent
Generative CA	LLM/generative model	Natural dialogue	Safety and validation concerns [4]
Sentimentix	NLP + sentiment + context + risk routing	Modular, safety-aware	Prototype; requires validation

As summarized in Table I, Sentimentix combines established components rather than claiming that one model can solve mental-health interaction by itself. The design is intentionally modular so that the sentiment classifier or response generator can be replaced as better validated models become available.

PROPOSED METHODOLOGY

The proposed Sentimentix architecture consists of five functional layers: input acquisition, NLP preprocessing, sentiment/emotion inference, context and safety reasoning, and response generation. Fig. 1 presents the overall architecture. The modular design enables independent testing of each stage and reduces the risk of coupling clinical interpretation to a single prediction model.

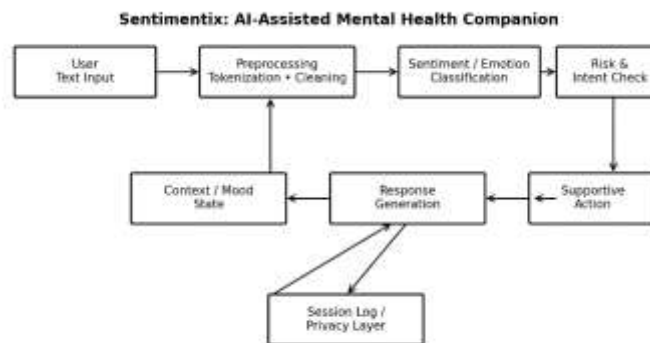


Fig. 1. Proposed architecture of the Sentimentix mental-health companion.

A. Input and Preprocessing: The user enters a natural-language message through the conversational interface. The text is normalized by removing unnecessary symbols, standardizing case where appropriate, and tokenizing the message. Additional preprocessing may include lemmatization, stop-word handling, spelling normalization, and removal of irrelevant artifacts. Preprocessing is configurable



because aggressive normalization can remove useful emotional cues such as repeated punctuation or capitalization.

B. Sentiment and Emotion Classification: The processed message is transformed into features suitable for a classification model. A baseline implementation can use TF-IDF with a classical classifier, while an advanced implementation can use an LSTM or transformer encoder. The classifier produces a sentiment label such as positive, neutral, or negative and may optionally produce finer emotion categories such as sadness, anger, fear, joy, or frustration. The model output is treated as an uncertain estimate rather than a clinical measurement.

C. Context and Mood State: Single-message classification may be misleading. Sentimentix therefore maintains a lightweight session-level context containing recent user messages, recent sentiment labels, and response history. A rolling state can be represented as a sequence $S_t = \{s_{(t-k)}, \dots, s_t\}$, where s_t is the current sentiment/emotion prediction. This state allows the system to distinguish a transient negative message from repeated negative interactions.

D. Risk and Intent Layer: Before final response generation, the system checks for intent categories such as general conversation, emotional support, coping request, persistent distress, and high-risk language. The risk layer should use conservative thresholds and explicit safety rules. It must not claim to diagnose a disorder or assess suicide risk with clinical certainty. When high-risk indicators are present, the system should prioritize immediate human or emergency support guidance appropriate to the deployment context.

E. Response Generation: For ordinary supportive interactions, Sentimentix selects or generates responses aligned with the detected emotional state. Responses should use empathetic but non-deceptive language, avoid overclaiming expertise, and encourage healthy coping or professional support when appropriate. Template-based responses improve predictability, while a generative model can increase conversational flexibility if accompanied by strong safety constraints.

TABLE II. MAJOR MODULES OF SENTIMENTIX

Module	Input	Output	Purpose
Preprocess or	Raw text	Clean tokens	Prepare NLP input
Sentiment model	Tokens/features	Polarity/emotion	Estimate affective state
Context manager	Recent dialogue	Session state	Preserve continuity
Risk/intent layer	Text + state	Routing class	Safety-aware response selection
Response engine	State + intent	Reply	Generate supportive interaction

Fig. 2 summarizes the operational workflow. The system begins a session, receives a user message, preprocesses the text, estimates sentiment/emotion, evaluates intent and potential risk, and then chooses between an ordinary supportive response and a safety-oriented pathway.

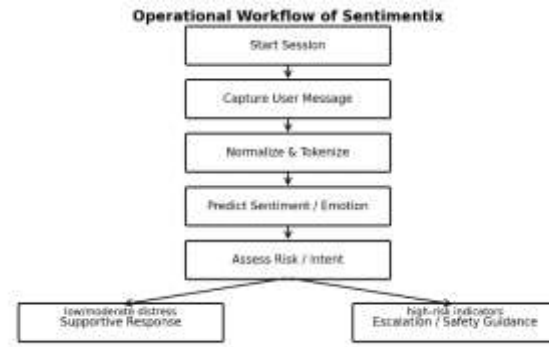


Fig. 2. Operational workflow of Sentimentix.

The response policy can be expressed as a simple decision function. Let e_t denote the estimated emotional state, c_t the conversational context, and r_t the risk/intent score. The response class y_t is selected as $y_t = f(e_t, c_t, r_t)$. When r_t is below the configured safety threshold, the system can select supportive conversational content conditioned on e_t and c_t . When r_t exceeds the threshold, the safety pathway is selected and the system provides appropriate guidance rather than attempting to manage a crisis autonomously.

A. Data and Model Strategy: A production implementation should use a dataset with clearly documented provenance, annotation guidelines, class definitions, and demographic considerations. Training and test sets must be separated to reduce leakage. Class imbalance should be examined because neutral or generic messages may dominate some datasets while high-risk expressions may be comparatively rare. Data augmentation can be considered cautiously, particularly for emotion classes, but synthetic examples should not replace representative real-world evaluation.

B. Model Evaluation: Sentiment classification can be evaluated using accuracy, precision, recall, macro-F1, and a confusion matrix. Macro-F1 is especially useful when classes are imbalanced. For the conversational layer, evaluation should include response relevance, empathy, coherence, safety, and refusal/redirect behavior. User-centered evaluation should use validated questionnaires and appropriate study designs rather than relying only on language-model metrics. Recent reviews emphasize that automated language metrics do not necessarily correspond to user well-being [4].

C. Privacy and Ethical Design: Mental-health conversations can contain highly sensitive personal information. Sentimentix should minimize data collection, avoid storing raw conversations unless necessary and consented to, encrypt retained information, and provide clear disclosure about the system's limitations. Where possible, local inference can reduce unnecessary transmission of personal text. Access control and retention policies should be defined before deployment.

D. Safety Constraints: Sentimentix should never present itself as a licensed therapist, make a clinical diagnosis from text alone, or imply that automated support is sufficient for severe mental-health emergencies. Safety messages should encourage contact with qualified professionals or appropriate local emergency resources when necessary. The 2026 literature continues to identify gaps in external validation and safety/governance reporting for LLM-based mental-health systems [4], reinforcing the need for conservative deployment.

E. Implementation Stack: A prototype can be implemented using Python, an NLP library such as scikit-learn or a transformer framework, a lightweight web interface, and a structured response/routing layer. A



database is optional for a privacy-first prototype; if persistence is introduced, only necessary metadata should be retained. The modular design also allows future replacement of the baseline classifier with a validated transformer model without redesigning the complete application.

RESULTS AND DISCUSSION

The present paper treats Sentimentix as a prototype research framework. Therefore, the results reported here describe functional and design-level outcomes that can be verified through scenario-based testing; no clinical efficacy, diagnostic accuracy, or therapeutic benefit is claimed without a controlled user study.

A. Functional Verification: Representative test scenarios can be constructed to cover positive, neutral, negative, anxious, frustrated, and potentially high-risk messages. For each scenario, the expected behavior is compared with the actual pipeline output: preprocessing completion, sentiment/emotion prediction, context update, intent/risk routing, and response generation. This verifies whether the modules operate in the intended sequence.

B. Sentiment-Aware Response Behavior: The central functional expectation is that messages with different emotional signals should not receive identical responses. A negative or distressed message should produce acknowledgment and supportive language, while a neutral informational message can receive a direct response. Repeated negative signals should update the session state and can increase the priority of supportive or professional-help suggestions.

C. Safety Routing: Safety testing should include messages that contain explicit or indirect high-risk indicators, ambiguous distress, and ordinary emotional complaints. The objective is not to prove clinical risk prediction but to verify that the system does not provide inappropriate reassurance or unsafe instructions when serious language is detected. This is particularly important because recent reviews show that safety safeguards are inconsistently reported across mental-health conversational-agent research [4], [8].

D. User Experience: A useful prototype evaluation should ask users to rate perceived empathy, relevance, clarity, response speed, privacy confidence, and overall usability. These measures should be collected separately from clinical outcomes. The literature indicates that engagement and the perceived human–AI relationship influence user experience [2]. Consequently, a technically correct sentiment label is insufficient if the resulting response is perceived as robotic, dismissive, or repetitive.

TABLE III. PROTOTYPE TEST SCENARIOS AND EXPECTED BEHAVIOR

Scenario	Detected State	Expected Response Strategy	Safety Priority
Positive update	Positive/joy	Encourage and reinforce	Low
Routine question	Neutral	Direct informative reply	Low
Sadness/loneliness	Negative/sad	Empathy + coping suggestion	Moderate
Repeated distress	Negative + persistent	Supportive response + professional help suggestion	Elevated
High-risk wording	Risk indicator	Safety-oriented escalation guidance	High



Table III provides a practical basis for functional verification. It should not be interpreted as a clinical decision table. In a full evaluation, the scenario set would be expanded and independently annotated by domain experts. Model performance should then be reported with confidence intervals and error analysis.

E. Discussion

The principal advantage of Sentimentix is the integration of emotional awareness with conversational context and safety routing. This addresses a limitation of simple FAQ-style chatbots, which may respond appropriately to the words in a message but ignore the user's affective state. The approach is also consistent with the literature showing increasing use of sentiment analysis, personalization, mood tracking, and empathic responses in mental-health conversational agents [2].

However, several limitations remain. Sentiment is not equivalent to mental-health status; sarcasm, cultural expression, code-switching, and ambiguous language can cause classification errors. A user can also express positive language while experiencing serious distress. In addition, a response that appears empathetic may still be clinically inappropriate. These limitations mean that Sentimentix should be positioned as a supportive companion rather than an autonomous therapeutic system.

CONCLUSION

This paper presented Sentimentix, a sentiment-aware mental health companion that combines NLP preprocessing, sentiment/emotion classification, conversational context, safety-oriented intent routing, and supportive response generation. The proposed architecture is modular and can be implemented with lightweight NLP components or extended with transformer-based models. Its principal design objective is to make conversational support more context-sensitive while explicitly separating technical emotion recognition from clinical diagnosis.

The literature indicates that conversational agents can provide accessible support and may produce beneficial short-term outcomes in selected settings, but evidence remains heterogeneous and long-term clinical effectiveness is not established for every system [1], [2], [6], [7]. Sentimentix therefore emphasizes conservative claims, privacy, transparency, and escalation rather than presenting automated conversation as a substitute for professional care. Functional scenario testing can verify the behavior of the prototype, but future work must include systematic model evaluation, expert review, usability studies, and ethically approved human-subject research before any clinical interpretation is made.

FUTURE SCOPE

Several extensions can strengthen Sentimentix. First, transformer-based language and emotion models can replace or complement baseline classifiers to improve contextual understanding. Second, multilingual and code-switched models can broaden accessibility, particularly for users who naturally mix languages. Third, longitudinal mood tracking can identify changes across sessions, provided that explicit consent and privacy-preserving storage are implemented.

Fourth, human-in-the-loop escalation can connect the system to trained counselors or designated support personnel when predefined safety conditions are met. Fifth, explainable sentiment predictions can show the system's confidence and uncertainty instead of presenting emotional labels as facts. Sixth, evaluation should move from laboratory scenarios toward prospective, ethically approved studies using validated psychological measures and independent safety assessment.

Finally, future versions may incorporate voice, multimodal interaction, personalized coping resources, and retrieval from vetted mental-health information sources. These enhancements should be introduced



only with rigorous privacy, security, bias, and safety testing. Recent systematic reviews emphasize that external validation, standardized outcomes, and governance remain major gaps in AI-based mental-health systems [4], [8]. Addressing these gaps will be essential for moving Sentimentix from an academic prototype toward responsible real-world deployment.

REFERENCES

- [1] Y. He et al., “Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials,” *Journal of Medical Internet Research*, vol. 25, e43862, 2023, doi: 10.2196/43862.
- [2] H. Li, R. Zhang, Y.-C. Lee, and R. E. Kraut, “Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being,” *npj Digital Medicine*, vol. 6, 236, 2023, doi: 10.1038/s41746-023-00979-5.
- [3] Y. M. Cho, S. Rai, L. Ungar, J. Sedoc, and S. C. Guntuku, “An Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives,” in *Proc. EMNLP*, 2023, pp. 10711–10728.
- [4] H. N. Cho et al., “Large Language Model-Based Chatbots and Agentic AI for Mental Health Counseling: Systematic Review of Methodologies, Evaluation Frameworks, and Ethical Safeguards,” *JMIR AI*, vol. 1, e80348, 2026, doi: 10.2196/80348.
- [5] “Integrating sentiment analysis to enhance mental health support chatbots,” in *Proc. 2nd Int. Conf. Computing and Data Analytics (ICCDa)*, IEEE, 2024, pp. 1–6, doi: 10.1109/ICCDa64887.2024.10867341.
- [6] H. Gaffney, W. Mansell, and S. Tai, “Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review,” *Journal of Medical Internet Research*, vol. 21, no. 4, e14166, 2019, doi: 10.2196/14166.
- [7] Y. Hang et al., “The effectiveness of CBT-based NLP-enabled AI conversational agents for mental health intervention: a systematic review and meta-analysis,” *npj Digital Medicine*, 2026, doi: 10.1038/s41746-026-02886-x.
- [8] J. Gong et al., “Evaluating Text-based Conversational Agents for Mental Health: A Systematic Review of Metrics, Methods and Usage Contexts,” *arXiv:2602.17669*, 2026.