Psychological Health Chatbot, Detecting and Assisting Patients in their Path to Recovery

Sadegh Jafari[1], Erfan Zare[1], Amirreza Vishteh[2],
Melikeh Mirzaei[3], Zahra Amiri[1], Simamohammad Parast[3], Sauleh Eetemadi[4]
[1]Iran University of Science and Technology, [2]Sharif University of Technology,
[3]Islamic Azad University, [4]University of Birmingham
{sadegh_jafari, zahra_amiri}@comp.iust.ac.ir,
e_zare@elec.iust.ac.ir, amirreza.vishteh@ce.sharif.edu,
{mirzaeimelike, simamohammadparast}@gmail.com,
s.eetemadi@bham.ac.uk

Abstract

Mental health disorders such as stress, anxiety, and depression are increasingly prevalent globally, yet access to care remains limited due to barriers like geographic isolation, financial constraints, and stigma. Conversational agents or chatbots have emerged as viable digital tools for personalized mental health support. This paper presents the development of a psychological health chatbot designed specifically for Persian-speaking individuals, offering a culturally sensitive tool for emotion detection and disorder identification. The chatbot integrates several advanced natural language processing (NLP) modules, leveraging the ArmanEmo dataset to identify emotions, assess psychological states, and ensure safe, appropriate responses. Our evaluation of various models, including ParsBERT and XLM-RoBERTa, demonstrates effective emotion detection with accuracy up to 75.39%. Additionally, the system incorporates a Large Language Model (LLM) to generate messages. This chatbot serves as a promising solution for addressing the accessibility gap in mental health care and provides a scalable, language-inclusive platform for psychological support.

1 Introduction

Mental health issues, such as stress, anxiety, and depression, are increasingly prevalent worldwide, affecting millions of individuals (Prince et al., 2007). Access to effective mental health services, however, remains limited due to barriers such as geographic location, financial constraints, and societal stigma (Javed et al., 2021).

This paper introduces a psychological health chatbot designed to assist individuals in managing their mental health. The chatbot's primary functions include detecting emotions, identifying potential mental health disorders, and ensuring the safety and appropriateness of its responses. The chatbot is specifically designed for the Persian language, filling a critical gap in mental health care for non-English speaking communities.

The proposed system integrates several modules: emotion detection, disorder identification, and language model validation, ensuring safe, supportive interactions. Using the ArmanEmo dataset, a Persian emotion detection dataset, and advanced NLP techniques, the chatbot offers personalized, culturally relevant mental health support (Mirzaee et al., 2022). The development and evaluation of this chatbot contribute to the growing field of AI-driven solutions for mental health care, offering a resource that is more accessible and language-inclusive.

2 Related Works

Artificial intelligence (AI) and machine learning have increasingly been applied to mental health diagnosis, leveraging data from social media and digital platforms for early detection and intervention. Sophisticated AI chatbots are now capable of providing real-time mental health support (Team Capacity, 2023). Research indicates that AI can provide an affordable supplementary approach to traditional therapies, effectively aiding in the reduction of depressive and anxiety symptoms (Kaywan et al., 2023). With an average satisfaction rating of 3.95 out of 5 (79%), user feedback demonstrates substantial satisfaction and engagement levels (Kaywan et al., 2023). A non-clinical randomized trial platform further underscores the efficacy of AI-driven computer-assisted cognitive-behavioral therapy (CCBT) in alleviating self-reported depression and anxiety symptoms among college students (Fulmer et al., 2018). A study examining the effectiveness of CBT-based smartphone applications with 28 participants utilized the Shim chatbot, a text-based smartphone app, to collect data over a two-week period. The findings highlighted positive user experiences and outcomes from interactions with the chatbot (Ly et al., 2017).

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 64–77
January 19, 2025. ©2025 Association for Computational Linguistics

Findings suggest that GPT is a highly effective tool for identifying psychological constructs within textual data across multiple languages. Compared to traditional methods like dictionary-based and fine-tuned machine learning models, GPT offers notable advantages: it performs consistently across languages and contexts, eliminates the need for training data, and operates with minimal coding through straightforward prompts (Rathje et al., 2024). GPT has demonstrated significantly enhanced accuracy in detecting annotated sentiments and discrete emotions, outperforming commonly used dictionary-based methods prevalent in psychology and social sciences (Jackson et al., 2022).

The World Health Organization (WHO) notes a growing global need for mental health services (World Health Organization, 2023), and machine learning offers scalable solutions to address this demand by analyzing large datasets for risk prediction. Reports from the Australian Bureau of Statistics (Australian Bureau of Statistics, 2021) and the U.S. Department of Health and Human Services (U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, 2018) underscore the increasing prevalence of mental health disorders, stressing the need for technological innovations. Machine learning and deep learning models have shown effectiveness in diagnosing mental health conditions from digital data. Iyortsuun et al. (Iyortsuun et al., 2023) review these techniques, finding deep learning methods particularly adept at identifying complex patterns, such as predicting suicidal tendencies from social media content (Wies et al., 2021). Challenges remain, particularly regarding stigma and self-stigma, which hinder help-seeking behavior (Clement et al., 2015; Oexle et al., 2017). Digital interventions, like AI chatbots, offer promise by providing anonymous support. However, ethical considerations must be addressed to align these technologies with human-centric values (Bryant, 2023; The Center for Humane Technology, 2023).

3 Methodology

Mental health issues, such as stress and anxiety, are increasingly common. Traditional therapies, while effective, are often inaccessible due to geographic or financial barriers. Digital solutions like conversational agents offer personalized mental health support. This study develops a conversational agent with emotion detection, disorder identification, and response safety evaluation to assist users in improving mental health. You can see the structure of this conversational agent in Figure 1. As illustrated in the figure, the system processes user messages through several steps. First, the input messages are analyzed using the Emotion Classifier, the Disorder Detection module, and the Message Validator. The Emotion Classifier identifies the emotions conveyed in the input text. The Disorder Detection module determines whether the user is experiencing stress. Simultaneously, the Message Validator assesses whether the user's message aligns with the chatbot's intended purpose. If the message is unrelated, the system provides a default response, notifying the user that their input is not relevant to the chatbot and cannot contribute to improving their emotional state.

For messages deemed relevant, the system considers the current input alongside previous messages, assigning weights to earlier messages based on their temporal proximity to the latest input. Using this contextual information, an answer is generated by a LLM. The generated response is then validated to ensure it is non-toxic and does not elicit negative emotions. If the response passes validation, it is presented to the user as the chatbot's reply.

4 Emotion Detection Module

This module identifies emotions in user messages based on six primary categories: sadness, hate, fear, anger, happiness, and surprise. An additional label, *other*, is included to account for emotions beyond these categories. By analyzing the input text, the module detects the user's emotional state, which is then utilized to generate optimal responses aimed at fostering calmness and improving emotional well-being. Further details regarding the module's implementation and performance are provided in Appendix A.

The six primary emotions are described as follows:

- **Sadness**: Sadness is a negative emotional state often linked to experiences of loss, hopelessness, or failure. It arises in response to distressing events and is associated with reduced interest in activities, low energy, and a desire for isolation (Beck, 1976).
- **Hate**: Hate is an intense and negative emotion characterized by feelings of hostility and

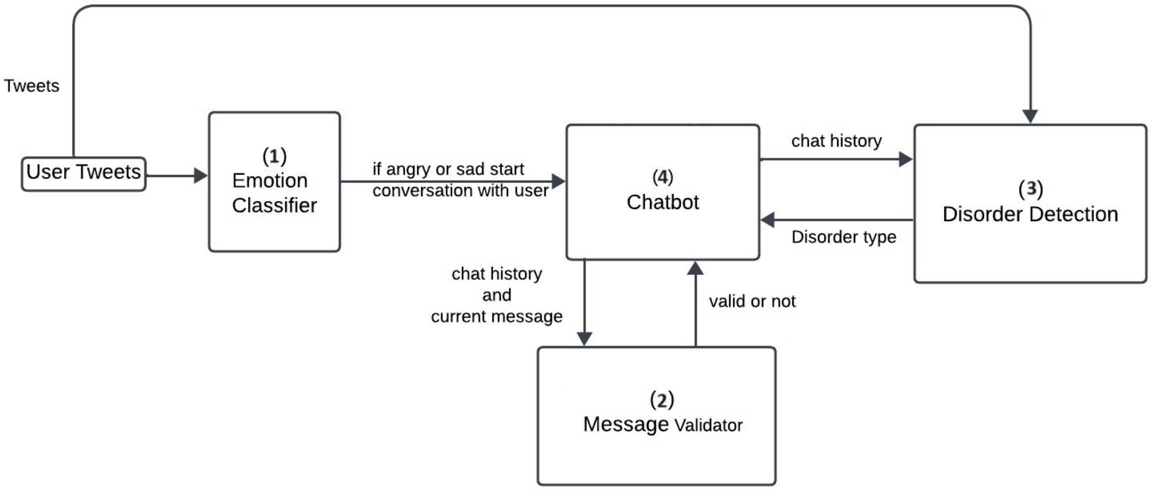

Figure 1: The structure of the mental health conversational agent. The system processes user messages through emotion classification, disorder detection, and message validation. Relevant messages are combined with contextual information to generate responses using a LLM. Responses are validated for non-toxicity before being delivered to the user.

disgust toward a specific target. It is associated with aggressive behaviors and hostility toward individuals or groups. Hate is recognized as one of the fundamental emotions in early theories of emotion (Izard, 1977).

- **Fear**: Fear is a natural response to real or perceived threats. It is marked by heightened alertness and readiness to confront or avoid danger. Physiological indicators, such as increased heart rate and sweating, are common markers of fear. The "fight or flight" theory highlights fear's role as a survival mechanism (Cannon, 1932).

- **Anger**: Anger typically emerges from provocations or frustrations and is often accompanied by a desire to confront the source of irritation. Behavioral indicators such as muscle tension and harsh vocal tones are associated with anger, which is seen as a natural regulatory response to challenges (Averill, 1982).

- **Happiness**: Happiness is a positive emotional state characterized by feelings of satisfaction, joy, and well-being. It is commonly expressed through smiling, social engagement, and other positive behaviors. Subjective measures of happiness demonstrate its validity as a distinct emotional construct (Lyubomirsky and Lepper, 1999).

- **Surprise**: Surprise is a brief reaction to unexpected events that often increases attention and focus. Nonverbal cues such as widened eyes and immediate verbal reactions are common indicators. Surprise is considered one of the primary emotions in studies of facial expressions (Ekman and Friesen, 1975).

4.1 LLM message validator

The module is designed to function as a filter, ensuring that messages generated by the LLM are neither toxic nor contain language that could evoke negative feelings in users. In this context, *toxic* language refers to expressions that are offensive, hateful, or emotionally harmful, including cyberbullying, harassment, and hate speech. Toxicity is inherently multi-dimensional and context-sensitive, requiring careful consideration of intent, language nuances, and social context. This aligns with the definition proposed by Sheth et al. (2021)., who emphasize the need for psychological and social theories to define toxicity while addressing ambiguities across its dimensions through explicit knowledge in computational models.

The six categories of toxicity used in this work are defined as follows (Al-Saffar et al., 2021):

- **Toxic**: General harmful, rude, or disrespectful comments.
- **Severe-Toxic**: A more extreme form of toxicity, often involving intense or persistent offensive language.
- **Obscene**: Comments containing vulgar or inappropriate language.
- **Threat**: Comments containing expressions of intent to harm others.
- **Insult**: Comments meant to demean or belittle someone.
- **Identity-Hate**: Comments targeting individuals or groups based on their identity, such as race, religion, gender, or ethnicity.

The performance metrics and detailed descriptions of the module are provided in Appendix B for further reference.

4.2 Users message validator

The goal of this section is to assess the relevance of user messages to psychology-related topics. Considering the diversity of users and the wide range of discussion topics, a data-driven approach was adopted for model design. To train the model, a dataset of user messages with the system was collected. This dataset included 1,025 messages, meticulously labeled by human experts into two categories: "psychology-related" and "non-psychology-related." The labeling process involved careful evaluation of each message's content based on criteria such as topic, tone, and the use of psychological terms or concepts.

As shown in Table 1, the dataset includes examples of messages, their translations, and assigned labels, which illustrate the distinction between "psychology-related" and "non-psychology-related" categories. The performance metrics and detailed descriptions of the module are provided in Appendix C for further reference.

4.3 Stress Detection

Based on Hans Selye's theory (Selye, 1956), stress is defined as a nonspecific response of the body to any demand or change, manifesting in three stages: alarm, resistance, and exhaustion. In the alarm stage, the body quickly responds to a challenge; during the resistance stage, it actively confronts the threat, and if stress persists, it enters the exhaustion stage, which can lead to physical and psychological issues.

Richard Lazarus and Susan Folkman (Lazarus and Folkman, 1984) define stress from a cognitive perspective as the result of an individual's mental appraisal of a situation and the available resources to cope with it. According to their theory, stress occurs when an individual perceives a situation as a threat or challenge that exceeds their coping abilities.

The performance metrics and detailed descriptions of the module are provided in Appendix D for further reference.

4.4 Content Generator

A LLM and three classification models are used to detect stress disorders, recognize user emotions, and evaluate chatbot responses to prevent inappropriate or toxic replies. The chatbot algorithm analyzes conversation history, calculates the weighted average of emotions and psychological disorders, and generates a short and friendly response in Persian. The chatbot uses emojis and informal language to create a more personable response without directly mentioning the user's stress or emotions. This chatbot has been used by around 190 people, who independently engaged with it since its development and the distribution of its link on LinkedIn by community members, and approximately 2,000 messages have been exchanged with it.

The psychological chatbot algorithm is designed to provide personalized and friendly responses to users. Its functioning can be broken down into the following steps:

- **Input and Output:** The algorithm has two main inputs:
  - `chat_history`: The conversation history between the user and the chatbot.
  - `window_size`: Defines how many messages from the conversation history should be considered.
  - `input_text`: The new message entered by the user.

  The output is a response generated by the AI, which is sent to the user.
- **Adjusting the Conversation History:** First, if the `window_size` is specified, the algorithm

Message (Original)	Translation (English)	Tag
امروز اصلا حالم خوب نیست. فکر می‌کنم همه ازم متنفرن.	I am not feeling well at all today. I think everyone hates me.	Related
برنامه نویسی بلدی؟	Do you know programming?	Not-Related

Table 1: Sample Messages with Translations and Labels

Algorithm 1 Generate AI Response for Psychological Chatbot

```
Input: chat_history, window_size, input_text
Output: AI response answer
if window_size then
    chat_history ← chat_history[:window_size]
end if
messages ← chat_history
emotion ← calculate_weighted_average(chat_history, 'emotion')
disorder ← calculate_weighted_average(chat_history, 'disorder')
Create prompt with context and user data as follows:

        The previous messages are the chat history between a patient and a psychologist. Suppose you are a professional
        psychologist. Based on the following information, respond to the patient with a short message. (Prevent to say 'Hi'
        in each message. And only speak in Persian)

        Emotional status: {emotion}

        Mental disorder status: {disorder}

        Patient message: {input_text}

        Speak more sincerely and informally, and use emojis to create a friendlier tone. Avoid mentioning the user's stress
        or emotion levels directly, and don't discuss them. Just be aware of these levels to respond appropriately.

messages.add({"role": "user", "content": prompt})
response ← openai.ChatCompletion.create(
        model = "gpt-4o-mini-2024-07-18",
        messages = messages
)
return response
```

limits the conversation history to the number of messages defined by `window_size`. This helps focus on recent messages to provide a more relevant response.

- **Calculating Emotions and Mental Status:** The algorithm then uses the `calculate_weighted_average` functions to calculate the weighted average of emotions (`emotion`) and mental disorder status (`disorder`) based on the messages in the conversation history. These values reflect the user's emotional and mental state throughout the conversation and are used to adjust the chatbot's response.

- **Creating a Prompt for the Model:** Using the calculated information (emotions and mental disorder status), the algorithm generates a prompt containing instructions for the model. This prompt directs the model to respond like a professional psychologist, focusing on the conversation without directly referring to the user's stress or emotional levels.

- **Adding New Message to Conversation History:** The user-generated message is added as the most recent entry to the list of messages.

- **Generating a Response with the GPT Model:** Finally, the algorithm uses the GPT model `gpt-4o-mini-2024-07-18` to generate a response. This model works with the input messages (`messages`) and provides a response based on the prompt and conversation history.

- **Returning the Response:** The algorithm returns the generated response, which is then displayed to the user.

This method helps the chatbot respond appropriately while considering the user's mental and emotional state, aiming to maintain a friendly and informal communication style.

4.5 User satisfaction

The user satisfaction form includes a series of questions, aimed at enabling participants to evaluate the quality and user experience of their interaction. Participants are asked to rate aspects such as the ease of understanding and responding to the chatbot; the resemblance of the experience to a psychiatric session in terms of time commitment; the effectiveness of text messaging compared to speaking with a psychiatrist; the efficacy of the question sequence in assessing depression levels; and the likelihood of recommending the interaction to friends and family in Iran. The form concludes with an open-ended question that allows participants to provide additional comments. These open-ended responses will be incorporated into future training phases.

By analyzing satisfaction rates and feedback, improvements will continue to be made to enhance interactivity and encouragement for future participants.

5 Results and Evaluations

The PHQ-9 is known for its unidimensional structure, solid validity, and reliability, and is regarded as a useful and effective tool in epidemiological and research contexts. Based on prior studies and the current data, it is suggested that the PHQ-9 may also be applicable in other contexts within the studied population, though further confirmation is needed.(Dadfar et al., 2018) The PHQ-9 is a self-administered scale used for screening, assessing, and monitoring depression severity.(Kroenke et al., 2003)

This scale consists of nine items that reflect symptoms over the past two weeks, with one item evaluating functional impairment (Association et al., 2015). Each item is scored on a 4-point Likert scale, ranging from 0 to 3: “not at all” (0), “several days” (1), “more than half the days” (2), and “nearly every day” (3). The total score on the PHQ-9, summing all nine items, ranges from 0 to 27. A score of ≥ 15 is classified as major depression, while a score of ≥ 20 indicates severe major depression. The diagnostic validity of the PHQ-9 for major depressive disorder (MDD) has been confirmed through studies in eight primary care settings and seven obstetric clinics (Kroenke et al., 2001).

Various versions of the PHQ-9 suggest different cut-off points, ranging from ≥ 9 to ≥ 13, with sensitivity levels between 73.8% and 77.5%, and specificity from 76.2% to 97%.(Santos et al., 2013; Khamseh et al., 2011)

The experimental procedure was conducted in three phases: before the initial interaction with the chatbot, after one week, and finally, at the conclusion of the 14-day period. Throughout this timeframe, users were required to engage in daily interactions with the chatbot.

A total of 14 participants were recruited for the experiment. Considering that participants were allowed to withdraw at any stage of the experiment (based on signing the consent form), one participant withdrew due to the sudden passing of their niece, two participants withdrew due to a lack of time, and three participants withdrew because the experiment was uninteresting or unattractive to them. Ultimately, the experiment was successfully completed by 9 participants. The detailed user information is presented in Table 2. During the experiment, emotions and stress levels were monitored and documented through the chatbot’s integrated modules.

Upon completion of the experimental period, an extensive analysis of the collected data was undertaken. Insights derived from user conversations, alongside emotion and stress data, yielded several notable observations. Firstly, as depicted in Figure 2, users exhibited higher stress levels at the beginning of the week, which gradually decreased midweek, only to rise again towards the end of the week and the start of a new one. Additionally, in relation to users’ emotional responses during interactions with the chatbot, it is evident from Figure 3 that participants predominantly expressed feelings of sadness, followed by happiness.

Moreover, as shown in Figure 4, 7 out of the 9 participants exhibited signs of improvement by the end of the experiment. However, 2 participants, identified by IDs 171 and 181, did not show signs of recovery, as indicated by their test results. A further review of these cases suggests that the chatbot may not be effective in providing immediate assistance for users suffering from severe depression. For such individuals, professional psychological intervention and treatment are recommended.

Gender	Location	Age	User ID
male	Zanjan	27	166
male	Kashan	24	171
female	Mashhad	23	172
male	Tehran	24	175
female	Tehran	16	179
male	Tehran	20	181
female	Tehran	47	187
male	Tehran	30	189
female	Yazd	22	191

Table 2: Table showing gender, location, age, and user ID.

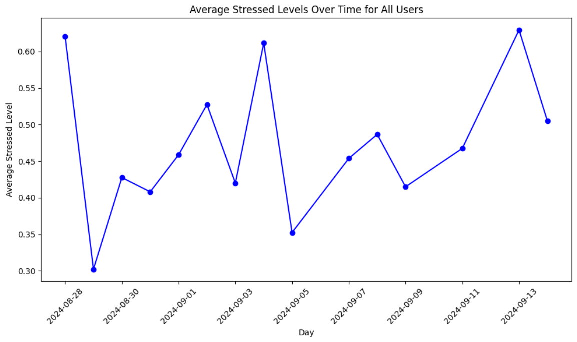

Figure 2: Average stress levels of all volunteer users over the two-week experiment

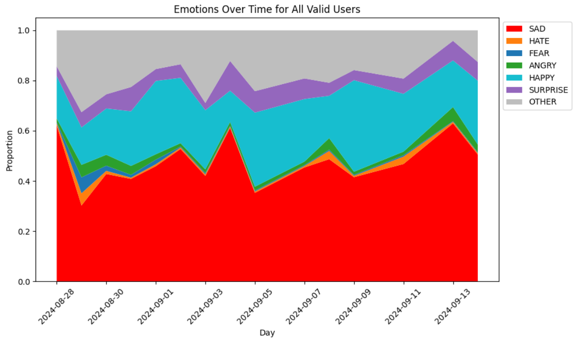

Figure 3: Average emotional responses of all volunteer users during chatbot interactions.

6 Conclusion and Future Works

In this section, we discuss our conclusions and the future work for this chatbot

6.1 Conclusion

The evaluation of the psychological chatbot demonstrated that it effectively facilitated natural and smooth interactions, offering valuable emotional feedback and responses aligned with cognitive-behavioral therapy principles. Users reported varying levels of satisfaction based on their initial mental health status, with those exhibiting higher levels of psychological distress showing less satisfaction. Despite these challenges, the chatbot successfully provided emotional reflections and relevant psychological techniques, contributing to improvements in users' anxiety and depression levels.

The chatbot's responses were generally accurate and addressed users' psychological issues, although its effectiveness varied. The analysis conducted by a licensed psychologist registered with the Iranian Psychological Association indicated that, while the chatbot adhered to cognitive-behavioral standards, it diverged from existential and Rogerian methods, which emphasize Socratic dialogue over structured techniques. User experiences were acceptable, with the chatbot meeting key criteria such as relevant responses, emotional reflection, and maintaining a coherent interaction memory.

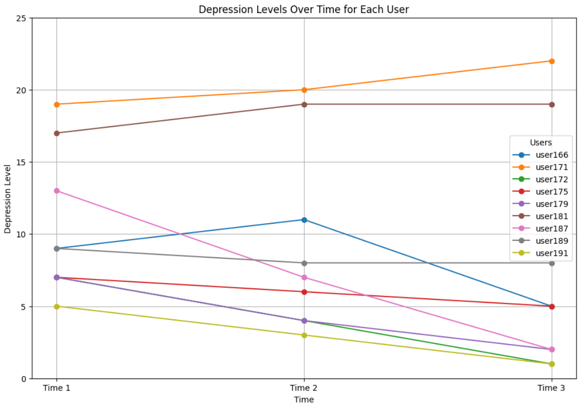

Figure 4: Results of the PHQ-9 questionnaire for all users.

A key strength of the system is its use of XLM-RoBERTa as the pre-trained model for multilingual capabilities, and ChatGPT-4.0 Mini, a multimodal model, enabling emotion detection and disorder identification to generalize effectively to other languages that use the Arabic script. This design extends the system's scope beyond Persian, making it applicable to other low-resource languages, enhancing its usability in diverse linguistic contexts.

However, the chatbot has limitations, including repetitive handling of some emotions and challenges in managing user anger. To address these issues and enhance the chatbot's capabilities, several improvements are suggested. These include recommending self-help resources, implementing user follow-up features, and configuring therapy sessions with specific protocols.

6.2 Future Work

Future developments should focus on improving the chatbot's performance by closely simulating expert psychologists' approaches and enhancing the system's ability to understand and respond to user emotions. Implementing a system for building user profiles and using past interaction data to tailor responses could significantly improve the chatbot's effectiveness. Adopting advanced techniques such as Retrieval-Augmented Generation (RAG) can enhance response relevance by leveraging historical conversation data.

To further advance the chatbot, expanding data collection efforts and improving data quality are essential. Collaboration with counseling centers and psychologists could provide valuable insights and data for refining the system. Adding voice communication capabilities would not only increase user engagement but also enhance comfort by offering voice responses and transcription services. These steps, along with ongoing refinement of models and protocols, will help bridge the gap between the chatbot and traditional psychological therapies, ultimately leading to a more effective and user-friendly tool.

7 Limitations

Despite the promising outcomes observed in the chatbot's performance, several limitations should be acknowledged. One major constraint is the lack of suitable hardware resources, particularly GPUs, which has hindered the development and fine-tuning of a custom language model tailored for the mental health domain. Due to this limitation, we were compelled to rely on OpenAI's pre-trained models, which may not fully capture the nuances of mental health dialogue, especially in handling complex psychological states such as anger or deeper existential concerns. The reliance on external models also introduces challenges in achieving complete control over the model's behavior, potentially affecting the precision of psychological techniques used by the chatbot.

Another significant limitation lies in the evaluation process. Psychological interactions are inherently dynamic and personal, making it difficult to create repeatable experiments with consistent results. User experiences and responses vary across different sessions, even with the same user, due to changes in mental state, environmental factors, and timing. Consequently, establishing a controlled experiment with identical conditions for all users proved to be a challenge. This variability in user interaction presents difficulties in benchmarking the chatbot's performance consistently, as real-world psychological factors introduce noise that is hard to quantify or replicate in a laboratory setting. These limitations highlight the need for further improvements in both model customization and experimental design to enhance the chatbot's reliability and overall effectiveness.

Ethics Statement

This study focuses on human behavior and moods, with ethical considerations addressed through strict adherence to established guidelines to ensure the validity of the methods and approaches employed. Particular attention is given to safeguarding participants' privacy. Access to raw data is restricted exclusively to the research team, ensuring that unauthorized individuals cannot gain access. Participants are assured that all data remains anonymous to protect their privacy, and informed consent was obtained for their participation in this evaluation for educational purposes.

The development and deployment of a text-based empathetic chatbot also involve significant ethical considerations. Key concerns include protecting user data privacy, particularly emotional data, and implementing strict data protection measures to prevent misuse. It is emphasized that the chatbot is not a substitute for professional psychological or medical advice. The project is guided by the principle of beneficence, aiming to enhance user well-being and minimize harm. Additionally, the chatbot's development adheres to ethical standards of fairness, non-discrimination, and bias prevention.

References

Abrar Al-Saffar, Robert Khoury, Rana Zaki, et al. 2021. Social media toxicity classification using deep learning: Real-world application uk brexit. *Electronics*, 10(11):1332.

American Psychological Association et al. 2015. Patient health questionnaire (phq-9 & phq-2) construct: depressive symptoms. *Washington: APA*.

Australian Bureau of Statistics. 2021. National study of mental health and wellbeing. Available online: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release (accessed on 19 August 2023).

James R. Averill. 1982. *Anger and Aggression: An Essay on Emotion*. Springer, New York.

Aaron T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York.

A. Bryant. 2023. Ai chatbots: Threat or opportunity? *Informatics*, 10:49.

Walter B. Cannon. 1932. *The Wisdom of the Body*. W.W. Norton & Company, New York.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

S. Clement, O. Schauman, T. Graham, F. Maggioni, S. Evans-Lacko, N. Bezborodovs, C. Morgan, N. Rüsch, J.S.L. Brown, and G. Thornicroft. 2015. What is the impact of mental health-related stigma on help-seeking? a systematic review of quantitative and qualitative studies. *Psychological Medicine*, 45:11–27.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 100–110.

Mahboubeh Dadfar, Zornitsa Kalibatseva, and David Lester. 2018. Reliability and validity of the farsi version of the patient health questionnaire-9 (phq-9) with iranian psychiatric outpatients. *Trends in psychiatry and psychotherapy*, 40(2):144–151.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman and Wallace V. Friesen. 1975. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Prentice-Hall, Englewood Cliffs, NJ.

Liu et al. 2019. Roberta: A robustly optimized bert pretraining approach (fine-tuned on ud pos28 dataset). *arXiv preprint arXiv:1907.11692*.

Ibrahim Ezzat. 2020. Deep-translator.

Mehdi Farahani, Marzieh Ahmadi, Ehsan Kamalloo, Niloofar Safi Samghabadi, and Sabine Karimi. 2021a. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021b. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Russell Fulmer, Angela Joerin, Breanna Gentile, Lysanne Lakerink, Michiel Rauws, et al. 2018. Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, 5(4):e9782.

N.K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant. 2023. A review of machine learning and deep learning approaches on mental health diagnosis. *Healthcare*, 11:285.

Carroll E. Izard. 1977. *Human Emotions*. Springer, New York.

Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A Lindquist. 2022. From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3):805–826.

Afzal Javed, Cheng Lee, Hazli Zakaria, Robert D Buenaventura, Marcelo Cetkovich-Bakmas, Kalil Duailibi, Bernardo Ng, Hisham Ramy, Gautam Saha, Shams Arifeen, et al. 2021. Reducing the stigma of mental health disorders with a focus on low-and middle-income countries. *Asian journal of psychiatry*, 58:102601.

Jigsaw and Google. 2018. Jigsaw toxic comment classification challenge.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. In *arXiv preprint arXiv:1607.01759*.

Payam Kaywan, Khandakar Ahmed, Ayman Ibaida, Yuan Miao, and Bruce Gu. 2023. Early detection of depression using a conversational ai bot: A non-clinical trial. *Plos one*, 18(2):e0279743.

Mohammad E Khamseh, Hamid R Baradaran, Anna Javanbakht, Maryam Mirghorbani, Zahra Yadollahi, and Mojtaba Malek. 2011. Comparison of the ces-d and phq-9 depression scales in people with type 2 diabetes in tehran, iran. *BMC psychiatry*, 11:1–6.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2003. The patient health questionnaire-2: validity of a two-item depression screener. *Medical care*, 41(11):1284–1292.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Richard S. Lazarus and Susan Folkman. 1984. *Stress, Appraisal, and Coping*. Springer Publishing, New York.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. 2017. A fully automated conversational agent for promoting mental well-being: a pilot rct using mixed methods. *Internet interventions*, 10:39–46.

Sonja Lyubomirsky and Heidi S. Lepper. 1999. A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, 46(2):137–155.

H. Mirzaee, J. Peymanfard, H. Moshtaghin, and H. Zeinali. 2022. Armanemo: A persian dataset for text-based emotion detection. Available at: https://arxiv.org/abs/2209.14585.

N. Oexle, M. Müller, W. Kawohl, Z. Xu, S. Viering, C. Wyss, S. Vetter, and N. Rüsch. 2017. Self-stigma as a barrier to recovery: A longitudinal study. *European Archives of Psychiatry and Clinical Neuroscience*, 268:209–212.

Hossein Poostchi and Ali Zarei. 2016. Hazm: Python library for digesting persian text.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. 2007. No health without mental health. *The lancet*, 370(9590):859–877.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Iná S Santos, Beatriz Franck Tavares, Tiago N Munhoz, Laura Sigaran Pio de Almeida, Nathália Tessele Barreto da Silva, Bernardo Dias Tams, André Machado Patella, and Alicia Matijasevich. 2013. Sensitivity and specificity of the patient health questionnaire-9 (phq-9) among adults from the general population. *Cadernos de saude publica*, 29:1533–1543.

Hans Selye. 1956. *The Stress of Life*. McGraw-Hill, New York.

Amit Sheth, Valerie L. Shalin, and Ugur Kursuncu. 2021. Defining and detecting toxicity on social media: Context and knowledge are key. *arXiv preprint arXiv:2104.10788*.

Team Capacity. 2023. The complete guide to ai chatbots: The future of ai and automation. Available online: https://capacity.com/learn/ai-chatbots/ (accessed on 19 August 2023).

The Center for Humane Technology. 2023. Align technology with humanity's best interests. Available online: https://www.humanetech.com/ (accessed on 19 August 2023).

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration. 2018. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health. Available online: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2018R2/NSDUHDetTabsSect8pe2018.htm#tab8-28a (accessed on 19 August 2023).

Sandra Bringay Waleed Ragheb, Jérôme Azé and Maximilien Servajean. 2019. Attention-based modeling for emotion detection and classification in textual conversations.

B. Wies, C. Landers, and M. Ienca. 2021. Digital mental health for young people: A scoping review of ethical promises and challenges. *Frontiers in Digital Health*, 3:697072.

World Health Organization. 2023. Mental health. Available online: https://www.who.int/health-topics/mental-health#tab=tab_1 (accessed on 19 August 2023).

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

A Emotion Detection

1.1 Dataset Overview

The *ArmanEmo* dataset is a Persian-language emotion detection dataset with over 7000 manually labeled sentences. The sentences were sourced from platforms such as Twitter, Instagram, and Digikala and are categorized into seven emotion labels: Anger, Fear, Happiness, Hatred, Sadness, Wonder, and Other (for emotions not covered by the main emotion labels).

The dataset has been split into training and testing sets and provided in TSV format. Transfer learning experiments have shown that *ArmanEmo* is better suited for emotion detection tasks compared to other older Persian datasets (Mirzaee et al., 2022). See Table 3 for details on the data sources used for ArmanEmo.

Source	**Persian Tweets**	**Instagram Comments**	**Digikala Comments**
Collection Period	2017-2018	2017-2018	2018
Raw Data	1.5M	1M	50K
Labeled for Manual Annotation	3.5K	3K	1K
Data for Automatic Annotation	4.5K	-	-

Table 3: Data sources for the ArmanEmo dataset, including collection periods, raw data size, and data labeled through both manual and automatic annotation processes.

The dataset has been split into training and testing sets and provided in TSV format. Transfer learning experiments have shown that *ArmanEmo* is better suited for emotion detection tasks compared to other older Persian datasets (Mirzaee et al., 2022).

1.2 Model Performance and Testing

Various models were tested on the *ArmanEmo* dataset. Below are the results of the key models:

1. **ParsBERT**: A version of BERT optimized for the Persian language. Achieved an accuracy of 63.8575 after 17 epochs (Farahani et al., 2021b).
2. **RoBERTa-Facebook**: An optimized version of BERT developed by Facebook, which achieved an accuracy of 63.1625 after 5 epochs (Liu et al., 2019).
3. **RoBERTa-Base-ft-UDPOS28**: A version of RoBERTa fine-tuned for part-of-speech tagging, achieving 62.033 accuracy after 5 epochs (et al., 2019).
4. **XLM-RoBERTa-Large**: A multilingual version of RoBERTa trained on data from over 100 languages. This model performed the best, showing superior generalization capabilities on the *ArmanEmo* dataset (Conneau et al., 2020).

1.3 Performance Comparison

Table 4 provides a summary of the precision, recall, and F1 scores for each model tested.

Model	**Precision**	**Recall**	**F1**
FastText (Joulin et al., 2016)	54.82	46.37	47.24
HAN (Yang et al., 2016)	49.56	44.12	45.10
RCNN (Lai et al., 2015)	50.53	48.11	47.95
RCNNVariant (Lai et al., 2015)	51.96	48.96	49.17
TextAttBiRNN (Waleed Ragheb and Servajean, 2019)	54.66	46.26	47.09
TextCNN (Kim, 2014)	58.66	51.09	51.47
ParsBERT (Farahani et al., 2021b)	67.10	65.56	65.74
XLM-Roberta-base (Conneau et al., 2020)	72.26	68.43	69.21
XLM-Roberta-large (Conneau et al., 2020)	**75.91**	**75.84**	**75.39**

Table 4: Comparison of Model Performance on ArmanEmo Dataset for emotion detection task(Precision, Recall and F1 metrics are macro).

B LLM message validator

In this module, the generated text by LLM is evaluated to ensure that no inappropriate content is included in the user-provided text. Given the importance of vocabulary and its impact on users' mental well-being, text evaluation and generating suitable content aimed at improving the user's state of mind are critical tasks.

2.1 Implementation

Since the chatbot operates in Persian, access to and use of a Persian language dataset was necessary. Due to the unavailability of an appropriate Persian dataset, an English-language dataset was used and translated using existing translation APIs, such as `deep-translator` (Ezzat, 2020). Consequently, the "Jigsaw Toxic Comment Classification" dataset (Jigsaw and Google, 2018) was utilized as a reference. This dataset contains 159,571 samples with six labels, including "toxic," "identity_hate," "insult," "threat," "obscene," and "severe_toxic." Since the dataset is multi-label, it allows for the possibility that a sample may have multiple labels. After translation, preprocessing was performed using the `hazm` library (Poostchi and Zarei, 2016), which includes operations such as removing extra spacing and reducing the repetition of consecutive words. Moreover, untranslated English words were removed using Unicode.

Given the data imbalance, as clearly shown in Table 5 as first two columns, where the distribution of the labels is presented, the need to improve the biased dataset was identified. To address this, a balanced subset was selected for each label. Due to hardware limitations during training, the dataset was reduced to 20,000 samples, with the distribution of labels shown in Table 5 in two balanced columns. Initially, the data was split into training and testing sets in a 4:1 ratio. Hyperparameters were determined manually using trial and error, and the final hyperparameters used for training the models are as follows: the number of training epochs was set to 3, with a per-device training batch size of 8 and an evaluation batch size of 16. The learning rate was adjusted to 2e-5, and a weight decay of 0.01 was applied. For optimization, the AdamW optimizer was employed. The different models were then trained and evaluated based on the test data. The results are presented in Table 6. According to the obtained results, the `xlm-roberta-large` (Conneau et al., 2019) model was selected as the final model used in the `message validator` module to evaluate the LLM-generated text. The detailed results of the final model's evaluation on the test data are presented in Table 7.

Model	Precision	Recall	F1-score
BART-base (Lewis et al., 2020)	92.66	88.81	90.54
BART-large (Lewis et al., 2020)	93.44	88.58	90.82
ELECTRA-base (Clark et al., 2020)	91.47	83.06	86.92
ParsBERT (Farahani et al., 2021a)	**93.93**	91.62	92.99
BERT (Devlin et al., 2018)	93.69	91.08	92.49
XLNet-base (Yang et al., 2019)	90.44	86.39	87.97
DistilRoBERTa-base (Sanh et al., 2019)	92.63	88.72	90.63
DistilBERT (Sanh et al., 2019)	93.88	91.84	92.80
XLM-RoBERTa-large (Conneau et al., 2019)	92.35	**93.95**	**93.43**

Table 6: The performance of the proposed model on custom data that explained in Table 5.

Label	Precision	Recall	F1-score
toxic	94.16	97.55	95.83
severe-toxic	97.42	88.39	92.68
obscene	89.59	88.55	89.07
threat	99.19	97.79	98.48
insult	86.87	89.00	87.92
identity-hate	96.48	92.84	94.63
Overall-accuracy	70.53		

Table 7: The result of XLM-Robetrta-large model on balanced dataset 5 for LLM Text Validation module.Due to the multilabel and multiclass structure of the dataset, there are cases where some labels are correctly identified while others are missed. This causes differences between the Precision and F1-Score values and the Overall Accuracy.

2.2 Challenges in Implementation

One of the main challenges was the absence of a suitable Persian dataset, which required the translation of another dataset. Due to the weaknesses in translator APIs, such as inaccuracies in translating slang, idiomatic expressions, and offensive terms, this led to unbalanced translations or the non-translation of some words. Additionally, the limited availability of multi-class datasets with clearly labeled instances for different types of offensive or inappropriate sentences restricted the implementation to a specific dataset.

Label	Absent	Present	Balanced Absent	Balanced Present
toxic	144,277	15,294	4,879	15,121
severe-toxic	157,976	1,595	13,369	6,631
obscene	151,122	8,449	10,741	9,259
threat	159,093	478	8,129	11,871
insult	151,711	7,877	10,403	9,597
identity-hate	158,166	1,405	6,495	13,505

Table 5: Number of record counts in base dataset with balanced format for each label in the dataset for LLM Text Validation Module (Jigsaw and Google, 2018).

C Users message validator

In light of the possibility of irrelevant conversations occurring between users and chatbots, the necessity of implementing a module to evaluate the relevance of user messages with the chatbot's purpose has been identified. A dataset comprising conversations between users and the chatbot was collected and labeled accordingly.

Subsequently, preprocessing was performed on the generated dataset using the `hazm` library. This process involved correcting typographical errors, addressing literary issues in the text, eliminating repetitive characters, and removing stopwords. The final dataset, based on the distribution of labels, is presented in Table 8.

Given the limited size of the dataset, the cross-validation method was employed to train the models. The dataset was divided into five parts, with each iteration using four parts for training and one part for validation. This process was repeated five times so that each part was tested as a validation set. The hyperparameters used for training were optimized for transformer-based models as follows: the number of training epochs was set to 7, with gradient accumulation steps of 2. The per-device training batch size was set to 4, while the evaluation batch size was set to 8. The learning rate was adjusted to 2e-5, and a weight decay of 0.01 was applied. The results of the selected models are presented in Table 9.

Based on the results, it was observed that the

ParsBERT model outperformed others and was thus selected as the baseline model. In cases where user conversations were deemed irrelevant to the chatbot's purpose, a static message is sent to the user, and the API call is prevented, guiding the conversation back on track to improve the user's experience.Table **??** shows that the chatbot ignores texts that are not related to its purpose.

3.1 Challenges

Due to the limited number of samples in the dataset, there was a risk of overfitting during model training, which was mitigated by utilizing cross-validation. Additionally, certain conversations contained non-Persian text, emojis, or punctuation, necessitating further preprocessing to ensure high-quality data for model training.

Label	**Count**
Not Related	524
Related	738

Table 8: Number of Samples for Each Label in a collected dataset from user conversations.

Model Name	**F1-Score**	**Precision**	**Recall**
ParsBERT (Farahani et al., 2021a)	**95.26**	96.80	93.86
distil-bert multilingual	94.78	93.48	96.19
bert (Devlin et al., 2018)	91.86	96.23	88.68
XLM-Roberta-base (Conneau et al., 2019)	92.70	93.97	91.56
bart-base (Lewis et al., 2019)	81.78	88.37	76.45
DeBERTA-base	84.59	87.16	82.37
BART-large	81.49	83.24	80.08
electra-base (Clark et al., 2020)	67.55	79.92	62.63
xlnet-base (Yang et al., 2019)	50.11	33.71	**97.82**
XLM-Roberta-large	18.88	11.72	65.81

Table 9: Performance of Evaluated Models on the Collected Dataset.

D Stress Detection

4.1 Dataset Description

The dataset used for stress detection was constructed using text-based social media articles from Reddit and Twitter, as described in the paper titled *"Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study"*. The datasets are publicly available[1].

Dataset Overview: We constructed four high-quality datasets using text articles from Reddit and Twitter. Each article is annotated with a binary class label where:

- `0`: Stress Negative article
- `1`: Stress Positive article

The annotation was performed using an automated DNN-based strategy outlined in the aforementioned study.

The four datasets are described as follows:

- **Reddit Title:** Consists of titles from articles collected from both stress and non-stress-related subreddits on Reddit.
- **Reddit Combi:** Combines the title and body text from articles collected from both stress and non-stress-related subreddits on Reddit into a single text sequence.
- **Twitter Full:** Contains stress and non-stress-related tweets collected from Twitter.
- **Twitter Non-Advert:** A denoised version of the Twitter Full dataset, excluding advertising content.

4.2 Model Architecture

We employed a sequential neural network model to detect social media text stress. The architecture of the model is as follows:

- **Embedding Layer**: The embedding layer is initialized with 40-dimensional word vectors and a maximum input sequence length of 20 tokens. This layer contains 160,000 parameters.
- **Bidirectional LSTM Layer**: A Bidirectional Long Short-Term Memory (LSTM) layer with 100 units in each direction, yielding an output of 200 units. This layer consists of 112,800 parameters.
- **Dropout Layer**: A dropout layer is added to reduce overfitting by randomly dropping neurons during training with a dropout rate of 50%.
- **Dense Output Layer**: A fully connected dense layer with a sigmoid activation function is used for binary classification (stress vs non-stress), adding 201 trainable parameters.

The model has a total of 273,001 trainable parameters and achieves an accuracy of 98% on the test set, as summarized in Table 10.

[1] https://www.kaggle.com/datasets/mexwell/stress-detection-from-social-media-articles

Class	Precision	Recall	F1-Score
0 (Non-Stress)	0.96	0.99	0.97
1 (Stress)	0.99	0.96	0.98
Accuracy	0.98		

Table 10: Model Performance on Stress Detection Task.

The macro and weighted averages for precision, recall, and F1-score are consistently high, indicating robust performance across both stress and non-stress classes.