



ORIGINAL RESEARCH

Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support

Olajumoke Taiwo | Baidaa Al-Bander

School of Computing, Keele University, Keele, UK

Correspondence

Baidaa Al-Bander.

Email: b.al-bander@keele.ac.uk**Funding information**

Keele University

Abstract

Mental health disorders have a global prevalence of 25%, according to the WHO, and this is exacerbated by factors such as stigma, geographical location, and a worldwide shortage of practitioners. Mental health chatbots have been developed to address these barriers, but these systems lack key features such as emotion recognition, personalisation, multilingual support, and ethical appropriateness. This paper introduces an innovative mental health support system that integrates BERT-based emotional distress detection with a psychological first aid (PFA)-generative pre-trained transformer (PFA-GPT) model, providing an emotion-aware PFA chatbot. The methodology leverages deep learning models, utilising bidirectional encoder representations from transformers (BERT) for emotional distress detection and fine-tuning GPT-3.5 on therapy transcripts for PFA chatbot development. The findings demonstrate BERT's superior accuracy (93%) for emotional distress detection compared to bidirectional long short-term memory. The multilingual PFA chatbot developed using the PFA-GPT model demonstrated superior BERT scores (exceeding 83%) and proficiently provided ethical PFA. A proof of concept has been developed to illustrate the integration of the emotional distress detection model with the novel generative conversational agent for PFA. This integrated approach holds significant potential in overcoming existing barriers to mental health support and has the potential to transform mental health support, offering timely and accessible care through AI-powered psychological interventions.

KEYWORDS

artificial intelligence, learning (artificial intelligence), machine learning

1 | INTRODUCTION

Artificial Intelligence (AI) and chatbots have become increasingly popular across various fields, including healthcare, sports, mental health, and nutrition. This trend showcases AI's versatility and potential to enhance personalised services and user experiences. In healthcare, AI chatbots are emerging as valuable tools for delivering information and support, with the potential to transform service delivery [1]. These chatbots can automate routine tasks and provide tailored recommendations,

making them beneficial across multiple sectors. In this context, **psychological first aid (PFA)** refers to the initial support provided to individuals in emotional distress, aiming to alleviate suffering and promote recovery. **Emotional distress** encompasses a range of negative emotional states that can impact mental well-being.

Recent studies compare AI models such as ChatGPT and Google's Bard for various applications. Waisberg et al. [1] note distinct strengths among these models, indicating that although tools like Gemini are promising in language comprehension,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Cognitive Computation and Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Shenzhen University.

more development is necessary for reliable clinical applications. Furthermore, AI's potential in academic writing and research emphasises the importance of ethical considerations and human oversight [2,3]. In sports, Washif et al. [4] explored ChatGPT's capability to generate personalised resistance training programs. Their findings suggest that although AI can create structured regimens, human expertise is essential for maximising effectiveness. Similarly, in nutrition, Dergaa et al. [5–7] assessed ChatGPT's ability to offer dietary advice through hypothetical scenarios with distinct health conditions to assess the reliability of ChatGPT's nutritional recommendations. The results showed that although ChatGPT can provide basic dietary plans, it often overlooks specific health conditions and realistic weight loss goals, underscoring the need for professional input.

The application of AI in mental health is evolving presenting both opportunities and challenges. Dergaa et al. [2, 3] evaluated ChatGPT's effectiveness in mental health assessments and interventions through simulated interactions, concluding that although it can deliver empathetic responses, its recommendations may lack the specificity needed for clinical use. The integration of AI in healthcare has also made strides in areas such as precision medicine and patient flow management [8]. With the worsening burden of mental illness since the COVID-19 pandemic [9], leveraging technology for mental health assessment and management has become critical. The rise of digital mental health, driven by AI advancements, is evident in the proliferation of mental health apps [10].

Mental health disorders are a major global public health concern, affecting a significant portion of the population and imposing substantial societal and healthcare burdens. The World Health Organization estimates that one in four people will experience a mental health condition in their lifetime [11]. The impact extends beyond individual suffering, contributing to social, economic, and healthcare challenges. Approximately 20% of adults will face mental disorders, worsened by stigma, geographical disparities, and a shortage of practitioners [12]. Recent advancements in AI and natural language processing (NLP) have generated interest in AI-powered mental health interventions, which have the potential to bridge treatment gaps and provide accessible, personalised support. Research has focused on using AI to predict mental health risks [13] but gaps remain in employing AI to enhance digitally administered psychotherapies, such as PFA and internet-based cognitive behavioural therapy. Current efforts aim to develop conversational interfaces and chatbots for delivering psychotherapeutic interventions and fostering engagement. Noteworthy examples include Woebot for CBT delivery [14] and Wysa for managing depression symptoms [15]. Artificial intelligence techniques also facilitate text-based messaging with human coaches and identify 'positive change' markers in the cognitive processes of those experiencing mental distress [16]. Smartphone applications like mindLAMP and BiAffect leverage AI to analyse user data and predict mental health episodes [17]. Despite the promising applications of AI chatbots, stigma persists particularly in academic research. Dergaa et al. [2, 3] examined the origins and impacts of this stigma and proposed

strategies for overcoming it, emphasising the need to view AI chatbots as facilitators of efficiency and innovation rather than threats.

The remainder of this paper is structured as follows: Section 2 includes a review of related works focussing on AI-powered mental health support technologies (MHSTs). Section 3 covers data collection, model development, and evaluation, while Section 4 delves into the experimental outcomes and the proposed models' performance. Section 5 discusses the findings, and finally, Section 6 concludes the presented research work.

2 | RELATED WORKS

AI-based MHSTs encompass diverse classes such as machine and deep learning, computer vision, and NLP. Computer vision detects emotions, machine learning tracks mental health symptoms, deep learning powers neural network-based chatbots, and NLP enables sentiment analysis and emotional support chatbots. Machine learning techniques showed promise in aiding the detection and diagnosis of mental health conditions. For instance, they have successfully been used to differentiate between children with ADHD and bipolar disorder using wearable sensor data [18]. These techniques have also been applied to predict responders to drug treatment for Parkinson's disease [19]. Machine learning models leveraged annotated training data to understand contextual nuances in emotional expression and had been used by Facebook to detect signs of self-injury in posts [20]. Another study achieved 90% accuracy in classifying early depression signs using NLP and machine learning reported by the authors of Ref. [21]. In the context of suicidal ideation, machine learning algorithms have been employed to predict suicide risk using electronic health records, with an 80% accuracy rate within 2 years and 84% within a week [22].

Artificial intelligence also leveraged digital gaming and smartphone applications for mental health interventions, where gaming has evolved from symptom tracking to comprehensive programs for psychiatric disorders. Additionally, virtual reality helped to enhance mental health assessment by providing immersive experiences, covering various aspects including attention, response inhibition, and symptom evaluation, and offering innovative clinical potential for specific psychiatric symptoms [23–25]. Natural language processing plays a crucial role in mental health by extracting insights from self-reported indicators of psychological distress [26]. With roots dating back to the 1960s, NLP has evolved, combined with machine learning to detect conditions such as depression through linguistic markers on social media [27]. Natural language processing also aided in predicting psychosis onset with high accuracy [28,29] and employed sentiment analysis to extract emotional nuances from text, supporting text-based chatbots for timely emotional distress assistance [30].

Recent advances in deep learning, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based language models (LLMs),

have the potential to revolutionise mental health support systems by enabling empathetic and personalised chatbots. Convolutional neural networks are deep learning models tailored for analysing visual data, capturing intricate patterns, and discerning features in images and videos. Mohapatra et al. [31] demonstrated CNNs' effectiveness in predicting stress and anxiety levels through vocal data, achieving accurate emotion classification. Additionally, Kang et al. [32] developed an ensemble algorithm with a modified CNN-LSTM architecture, achieving 98.3% accuracy in classifying mental stress states through electrocardiogram (ECG) signals, showcasing potential in various healthcare applications. Recurrent neural networks, particularly LSTM and GRU variants, are adept at analysing sequential data in mental health applications, excelling in capturing emotional context and temporal dependencies for symptom detection and prediction. With internal memory states maintained through feedback loops, RNNs effectively model long-term dependencies in text data, demonstrating superior performance in identifying self-perceived symptoms of depression in mental health support systems [33–36].

Large language models (LLMs), such as bidirectional encoder representations from transformers (BERT) variants such as RoBERTa and ALBERT, play a crucial and effective role in analysing text-based interactions for mental health detection. Siddik et al. [37] highlighted BERT's superiority in mental health chatbots, whereas domain-specific pre-trained models such as MentalBERT demonstrated excellence in classification tasks [38]. Generative pre-trained transformers, such as GPT-2 and DialoGPT, with transformer architectures and self-attention mechanisms, have been applied in mental health support, where DialoGPT showed promising results, and ChatGPT demonstrated potential in tasks such as stress and depression detection. Fine-tuning LLMs, such as ChatGPT, for specific mental health applications holds promise for achieving high classification performance [39–46]. Ethical considerations in AI-driven mental health support systems involve prioritising user privacy, ensuring transparency, and mitigating bias in AI-driven mental health support systems [33,47–49]. Cultural appropriateness is emphasised, with multilingual chatbots like ChatPal [50] and SERMO [51] highlighting the importance of linguistically and culturally appropriate chatbots in enhancing mental healthcare.

While AI-based methods have shown promise in early detection, understanding, and automated diagnosis of affective disorders, there is a research gap in leveraging AI to enhance digitally administered psychotherapies. Mental health issues are a growing global concern, compounded by barriers in accessing personalised PFA. These barriers include stigma, limited access to services due to geographic and economic factors, and a shortage of mental health professionals. The reliance on crisis intervention over early care and the financial burden of treatment exacerbate the challenges. Additionally, the absence of multilingual capabilities in existing mental health support systems further restricts access for those seeking help in languages other than the default.

To address the issues with existing AI-powered mental health support systems, this work aims to leverage deep

learning and NLP to create a robust and innovative AI-driven mental health support system, offering personalised and accessible PFA and support. The specific aims encompass developing a multilingual system with real-time emotional distress detection and investigating the impact of fine-tuning conversational LLMs such as GPT-3.5 using therapy transcript data for PFA. The main contributions of this research are outlined as follows:

1. Designing an emotion aware PFA chatbot application, leveraging the capabilities of a large language model to provide advanced mental health support.
2. Introducing the PFA-generative pre-trained transformer (Psychological First Aid-Generative Pre-Trained Transformer (PFA-GPT)) model to generate psychologically informed responses. This model significantly enhances the effectiveness of mental health conversational agents.
3. Integrating state-of-the-art BERT-based emotional distress detection seamlessly with the PFA-GPT model. This integration introduces a methodology for generative AI in mental health support via chatbots, contributing to the advancement of technology-assisted mental health interventions.

3 | MATERIALS AND METHODS

The methodology for developing the AI-powered mental health support system involves thorough data preparation and pre-processing, model development, and fine-tuning, with a focus on ethical considerations. The methodology employs a mixed methods approach, utilising qualitative methods for assessing emotional distress detection and the conversational agent's performance, and quantitative methods for evaluating generated responses, considering metrics like accuracy, precision, recall, and BERTScore for quantitative evaluation.

3.1 | Dataset

Our study adopts diverse datasets from various sources, including emotionally annotated conversations and therapy transcripts, with a focus on emotion-labelled data for training and evaluating two key components: an emotional distress detection model and a generative conversational model for PFA. Ethical principles, such as user privacy and data protection, were carefully observed throughout. The first dataset, introduced by Rashkin et al. [52], contains around 25,000 conversations rooted in emotional situations, with labels for various emotions like sadness, anxiety, and anger. This data was essential for fine-tuning the BERT model's ability to detect emotional distress accurately. The second dataset, introduced by Bertagnolli [53], includes high-quality therapist responses to real patients' mental health questions. We used this data to train the conversational model, enabling it to generate empathetic and context-aware responses aligned with PFA principles.

3.2 | Data pre-processing

Data preparation and pre-processing involved removing null columns, renaming columns for clarity, eliminating irrelevant information and characters, applying data cleaning techniques, and removing duplicate records. Tokenisation is a foundational step in preparing data for BERT, breaking input text into tokens, including sub-words, essential for numerical processing. BERT's tokenisation method uses word piece tokenisation, special tokens like '[CLS]' and '[SEP],' and techniques for padding and truncation to handle varying input lengths. Each token is mapped to a numerical representation within BERT's vocabulary, receiving an embedding vector that encodes its semantic meaning and serves as input for subsequent processing. The self-attention mechanism in BERT allows tokens to evaluate their contextual significance by considering cues from all other tokens, enhancing the model's ability to capture complex semantic relationships. Positional information is integrated through positional encodings, resulting in contextualised representations for the entire input text, which can be employed as feature vectors for downstream tasks like classification or regression. Data is typically converted into tensors, which are multidimensional arrays processed efficiently by hardware. These tensor datasets represent input text after tokenisation and embedding, serving as numerical inputs for model training and inference. In our TensorFlow implementation, tensor datasets were effectively employed as

standalone tensors, although they can also be used as data loaders for handling batches of data.

3.3 | System design and architecture

The developed mental health support system consists of two distinct components, the emotional distress prediction model, and the generative conversational model for PFA, as shown in Figure 1.

The developed system, accessible through a user-friendly web application shown in Figures 2 and 3, integrates two user input channels: the chatbot and mood journal providing dual input options. The elements of the proposed architecture for the mental health support system illustrated in Figure 1, are detailed as follows:

- **Chatbot User Interface:** a page on the web application that serves as the route through which the user communicates with the chatbot.
- **Chat Input:** the message that the user sends to the chatbot with the intention of getting a response that could potentially provide PFA.
- **Journal User Interface:** This feature of the web application is a dedicated section enabling users to create personal notes about their emotional state and provides access to PFA in case the notes indicate signs of emotional distress.

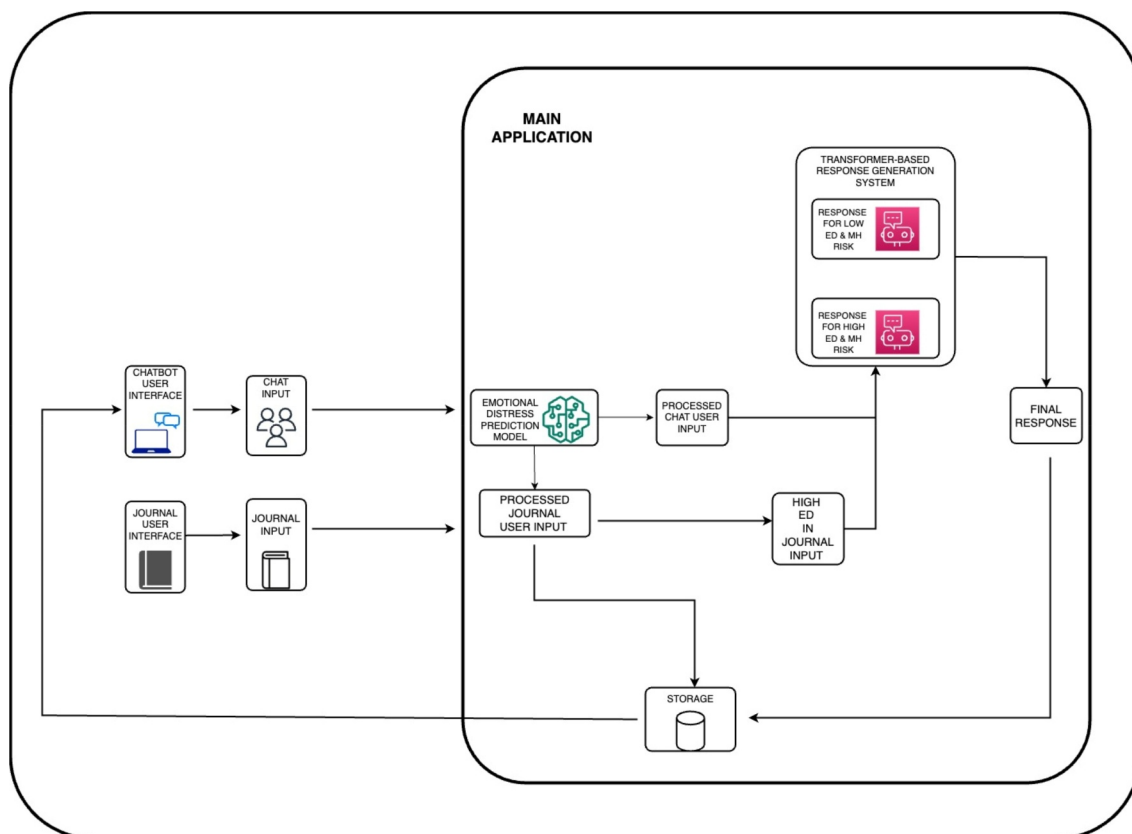
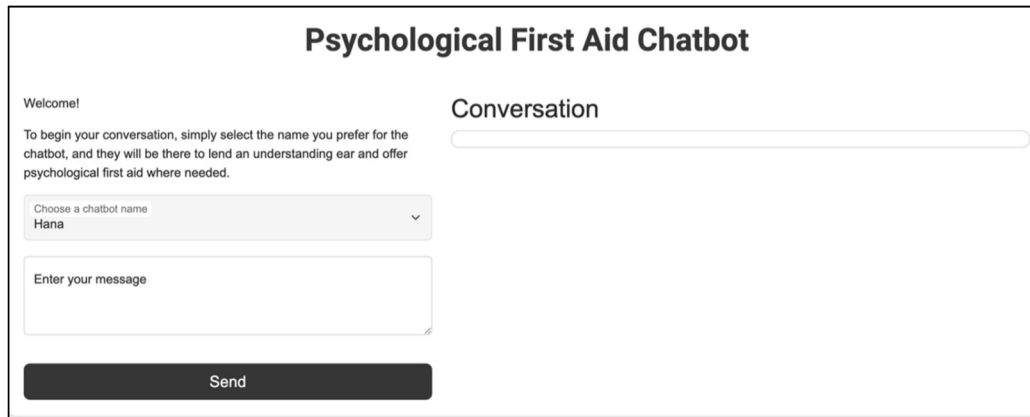


FIGURE 1 Proposed mental health support system architecture.

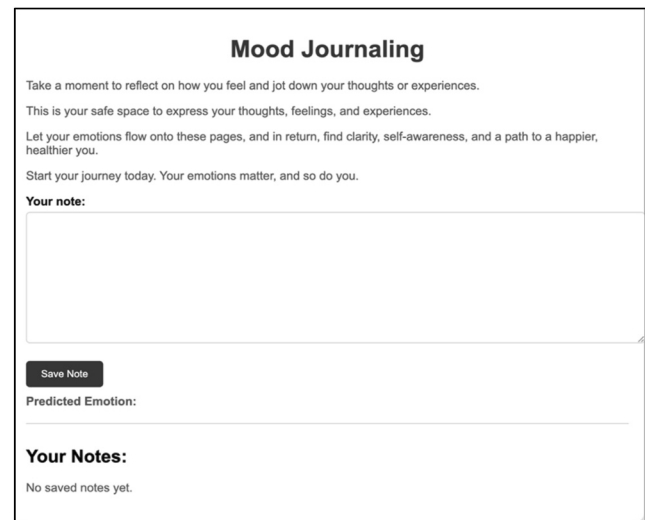


The interface is titled "Psychological First Aid Chatbot". It features a "Welcome!" section with instructions: "To begin your conversation, simply select the name you prefer for the chatbot, and they will be there to lend an understanding ear and offer psychological first aid where needed." Below this is a dropdown menu labeled "Choose a chatbot name" with "Hana" selected. There is a text input field labeled "Enter your message" and a "Send" button.

FIGURE 2 Psychological First Aid (PFA) chatbot interface.

- **Journal Input:** serves as input on the mood journal page of the web application, allowing users to record personal notes detailing their current emotional state.
- **Emotional Distress Prediction Model:** BERT-based model designed for classifying and predicting emotional states and distress levels based on inputs from chat conversations or journal entries.
- **Processed Chat User Input:** a combination of the user's chat input, the predicted outcome from the emotional distress prediction model, providing the predicted emotional state, and the emotional distress level within the chat input.
- **Transformer-based Response Generation System:** the proposed PFA-GPT model, which is a fine-tuned derivative of GPT-3.5. It generates contextually sensitive responses tailored to user's predicted emotional states and distress levels, providing PFA where necessary.
- **Processed Journal User Input:** a combination of the user's journal input, the predicted outcome from the emotional distress prediction model, providing the predicted emotional state, and the emotional distress level within the journal input.
- **Final response:** This is the response generated by PFA-GPT, it could provide empathetic, emotion-aware responses and PFA responses where necessary.
- **Storage:** The processed journal input and final response are stored for reference and to provide context.

Users engage with the chatbot by navigating to the designated chatbot page (Figure 2), selecting a preferred name for the chatbot, and inputting their message. Subsequently, the message undergoes analysis through an emotional distress prediction model to predict the user's emotional state and distress level. In case of a prediction indicating moderate or high distress, the response generation model—PFA-GPT—creates an emotionally sensitive response using prompts embedded with PFA principles. This response is then stored and presented to the user on the web page, enabling seamless continued interaction. Conversely, for low predicted distress levels, emotionally sensitive prompts generate supportive responses, indicating the chatbot's availability for listening and providing emotional support as needed.



The interface is titled "Mood Journaling". It includes a motivational message: "Take a moment to reflect on how you feel and jot down your thoughts or experiences. This is your safe space to express your thoughts, feelings, and experiences. Let your emotions flow onto these pages, and in return, find clarity, self-awareness, and a path to a happier, healthier you. Start your journey today. Your emotions matter, and so do you." Below this is a "Your note:" label followed by a large text input area. A "Save Note" button is positioned below the input area. Underneath the button is a "Predicted Emotion:" label followed by a text input field. At the bottom, there is a "Your Notes:" label and a message that says "No saved notes yet."

FIGURE 3 Mood journal interface.

Similarly, users could access the mood journal (Figure 3) through the dedicated mood journal page on the web application, where they can input their current emotions into the provided text box. Once the note is entered and saved, the system is prompted to analyse the journal input using the emotional distress prediction model to predict the emotional state and distress level in the journal input. If the analysis indicates low emotional distress, the note is saved with its emotional state information. Conversely, upon detecting moderate or high emotional distress, the system recommends engaging with the chatbot to potentially receive PFA support. Regardless of the user's choice to seek further assistance from the PFA chatbot, the note is saved with its emotional state information.

3.4 | Implementation of BERT-based emotional distress detection

The core of the architecture integrates BERT (TFBertForSequenceClassification) [54] a transformer-based model primarily using encoder layers. BERT's self-attention mechanisms enable

it to consider word dependencies and correlations, distinguishing it from traditional models by providing nuanced and accurate representations. Using a pre-trained BERT model eliminates the need for manual feature extraction, as it encodes contextual information, semantics, and word relationships. Bidirectional encoder representations from transformers comprises a stack of self-attention and feed-forward layers within its encoder, allowing it to focus on various aspects of input text regardless of word order. The model follows a pre-training and fine-tuning paradigm, where it learns contextualised word representations in the pre-training phase and is adapted to specific tasks, such as emotional distress detection, during fine-tuning. This process tailors the general language understanding of BERT to the task of identifying emotional distress levels.

3.5 | Development of Psychological First Aid conversational model and chatbot fine-tuning

In this study, we utilised OpenAI's GPT-3.5 [55], a transformer-based model with 175 billion parameters, to develop a therapy-focused conversational model. This model is particularly effective at autoregressive language modelling and generative text completion, utilising positional encodings to maintain word order and context throughout the sequences.

The fine-tuning process began with the preparation of a dataset formatted as a JSON Lines file, adhering to OpenAI's specifications. This dataset comprised user prompts paired with appropriate therapeutic responses. After the dataset was structured, it was uploaded to the OpenAI File API, enabling us to proceed with the fine-tuning. During the fine-tuning, we adjusted the training parameters to optimise the model for therapy-related tasks. This was a critical step to ensure that GPT-3.5 was effectively tailored to our specific application. When generating responses, we designed input prompts that included the user's initial input, emotional state, predicted distress levels, and instructions for relevant therapeutic strategies. This approach enabled the GPT-3.5 model to accurately comprehend the user's context and produce empathetic and PFA-like responses.

The model serves as the core component of our application, accessed through the OpenAI API. This API facilitates communication by transmitting user input and emotional context to the model, which then generates empathetic responses that are returned to the user interface, effectively completing the interaction.

3.6 | Evaluation metrics

The evaluation metrics collectively provide insights into the mental health support system's predictive capabilities and response quality. The metrics, accuracy, precision, recall, and F1 score provide a comprehensive evaluation of the BERT-based emotional distress prediction model. These metrics help to evaluate the ability of the model to accurately identify

emotional distress while minimising misclassifications. Accuracy offers a general performance overview, while precision minimises false positives by ensuring positive predictions are reliable. Recall reduces false negatives by ensuring actual distress cases are identified. The F1 score balances precision and recall, optimising the trade-off between avoiding false positives and false negatives. The confusion matrix also offers a detailed breakdown of predictions versus actual labels, enhancing understanding of the system's performance across various emotional states. These formulae are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1 Score} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

where TP, TN, FP, FN represent:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

For the GPT-3.5-based chatbot, bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), and BERTScore [41,42] are used to measure response quality and alignment with reference responses. Bilingual evaluation understudy scores gauge model-generated responses' similarity to human expert-provided reference responses, whereas ROUGE evaluates machine-generated text similarity against references text. Bilingual evaluation understudy assesses response similarity by comparing n - grams, whereas ROUGE evaluates generated content's quality based on n - grams. The BERTScore metric is a metric used to assess the quality of machine-generated text by comparing it to the reference text. BERTScore captures semantic meaning, ensuring fluency and content quality alignment. These formulae are defined as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (5)$$

where BP is the brevity penalty, N is the maximum n -gram order, w_n is the weight assigned to the n - gram precision, and P_n is the n - gram precision. Brevity penalty is defined as follows:

$$\text{BREVITY PENALTY} = \begin{cases} 1, & \text{if } c > r \\ e^{\left(1 - \frac{r}{c}\right)} & \text{if } c \leq r \end{cases} \quad (6)$$

ROUGE

$$= \frac{\text{Overlap of } N - \text{grams in generated and reference text}}{\text{Total number of } N - \text{grams in reference text}} \quad (7)$$

$$\text{BERTScore} = F_{(1)\text{BERT}} \text{ Score of token level overlap} \quad (8)$$

where:

$$F_{(1)\text{BERT}} \text{ Score} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

precision

$$= \frac{\sum_i \min(\text{reference token } count_i, \text{generated token } count_i)}{\sum_i \text{generated token } count_i} \quad (10)$$

recall

$$= \frac{\sum_i \min(\text{reference token } count_i, \text{generated token } count_i)}{\sum_i \text{reference token } count_i} \quad (11)$$

In the above formulae, i iterates over all tokens in the generated and reference text. The numerator counts the common tokens, and the denominator calculates the total tokens in either the generated or reference text.

4 | EXPERIMENTAL SETUP AND RESULTS

This section presents an evaluation of the proposed model's aptitude in offering mental health support by comparing the achieved outcomes with the baseline models using several performance metrics. The experimental setup for emotional distress detection using the BERT model [54] has involved fine-tuning on a diverse dataset [52], which is pre-processed to ensure uniformity. Assessment of emotional distress detection model performance encompassed accuracy, precision, recall, and F1-score metrics. The performance of the BERT model has been compared with a bidirectional long short-term memory (Bi-LSTM) model [56] for emotional distress detection on the same dataset. The architecture of the Bi-LSTM model capitalises on the LSTM's inherent ability to adeptly capture sequential relationships, making it well-suited for emotion classification tasks. The Bi-LSTM model's representation was further enhanced by the concatenation the LSTM outputs from the context and prompt sequences. This is carried out through a dense layer with rectified linear activation (ReLU), promoting non-linearity within the data. To prevent overfitting, a dropout layer randomly deactivated 50% of neurons. The final layer, activated through Softmax, generates probabilities corresponding to the array of emotion classes, yielding the model's predictions. Class weights are computed to

address class imbalances, and a learning rate scheduler dynamically adjusts learning rates to lower the learning rate by 10% after each epoch, promoting convergence. Early stopping strategy is adopted to prevent overfitting by monitoring validation loss. The hyperparameters to fine-tune the Bi-LSTM model are presented in Table 1.

The BERT tokeniser segments text into meaningful units and uses embeddings, including token, segment, and position embeddings. The model is then set up to classify emotions, and its parameters are fine-tuned through training iterations to minimise prediction errors and enhance accuracy. Batch processing is also used to facilitate efficient processing and parallelisation during training. The dataset is also shuffled to introduce randomness and enhance the learning process. Different hyperparameters are tested, and the specific combination that yielded the most optimal classification outcomes is provided in Table 2.

The fine-tuned BERT model, trained with an emotionally annotated dataset, showcased a notable accuracy of 93%. The visualised confusion matrix derived from the fine-tuned BERT model demonstrates a distinct diagonal pattern (See, Figure 4).

The precision, recall and F1 score metrics are also computed for the fine-tuned BERT and Bi-LSTM models, as shown in Table 3. The BERT model demonstrated higher level of correctness in positive prediction of emotions and achieved an impressive accuracy of 92.8% compared to the Bi-LSTM which had an accuracy of 81.6%. The computed F1-score for the fine-tuned BERT model stood at 0.9264, indicating a well-rounded performance, effectively harmonising precision and recall considerations.

For the conversational chatbot development, the GPT-3.5 and PFA-GPT models have been seamlessly integrated into the chatbot. Comprehensive testing is conducted across a spectrum of mental health scenarios, involving interactions designed to closely replicate authentic conversations. The evaluation criteria comprised aspects of personalisation, coherence, and relevance. The proposed PFA-generative pre-trained transformer (PFA-GPT) model was developed by fine-tuning the GPT-3.5 model. The fine-tuning process began with dataset preparation, data cleaning, and pre-processing. A subset of 200 data points was randomly selected and organised

TABLE 1 Hyperparameters to fine-tune Bi-LSTM model.

Hyperparameters	Batch size	Epoch	Bi-LSTM layers	Dropout layer
Values	64	20	2	0.5

Abbreviation: Bi-LSTM, Bidirectional Long Short-Term Memory.

TABLE 2 Hyperparameter combination for fine-tuning BERT model.

Hyperparameter	Learning rate	Batch size	Shuffle size	Optimiser
Value	2e-5	16	100	Adam

Abbreviations: BERT, Bidirectional Encoder Representations from Transformers.

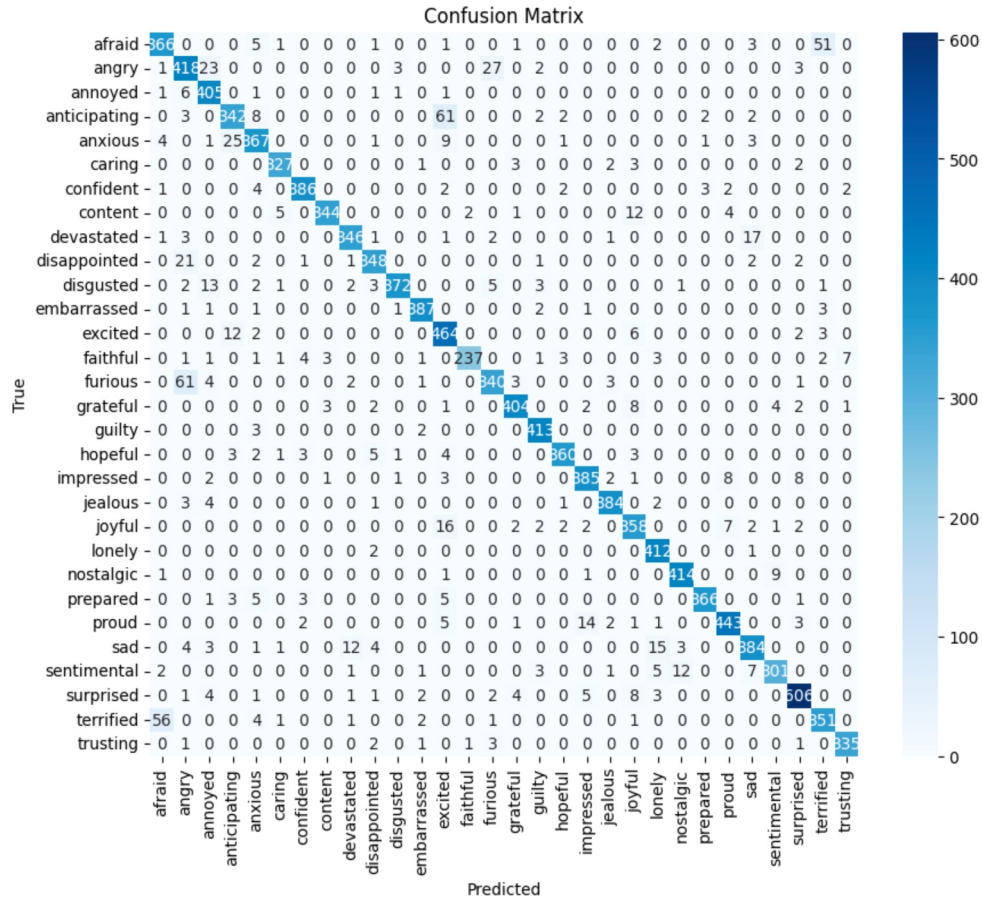


FIGURE 4 Multi-class classification of emotions using the fine-tuned BERT model. Abbreviation: BERT, Bidirectional Encoder Representations from Transformers.

TABLE 3 Comparison of BERT and Bi-LSTM for emotional distress detection using Dataset I.

Model	Accuracy	Precision	Recall	F1 score
Bi-LSTM	0.8155	0.8174	0.8155	0.8158
Fine-tuned BERT	0.9262	0.9280	0.9262	0.9264

Abbreviations: BERT, Bidirectional Encoder Representations from Transformers; Bi-LSTM, Bidirectional Long Short-Term Memory.

into a ‘therapy_transcripts.json’ file following OpenAI’s chat message format. This file was then uploaded via the OpenAI File API, and fine-tuning was initiated using the ‘gpt-3.5-turbo’ model, resulting in the creation of the novel conversational model, PFA-GPT. Psychological first aid-generative pre-trained transformer was designed for psychological support and integrated with the OpenAI API, connecting it to the chatbot. To enhance interaction, the fine-tuned BERT model predicted users’ emotional distress levels, allowing the chatbot to tailor its responses based on users’ emotional states for contextually aligned and sensitive interactions.

An evaluation dataset was created from the second dataset to assess the performance of the conversational models, including PFA-GPT and GPT-3.5, using metrics such as BLEU, ROUGE, and BERT scores. The evaluation scenarios contain 10 test user inputs that are used to generate responses

from both conversational agents. While BLEU and ROUGE scores provided lower values that underestimated response quality, BERTScore excelled in assessing semantic similarity and yielded scores of 83% and above for both models, as shown in Tables 4 and 5.

While BERTScore offers advantages in semantic similarity assessment, it may not fully account for fluency, coherence, and domain-specific nuances in dialogue. To address these limitations and evaluate relevance for PFA, qualitative assessment involved side-by-side comparisons of question–response pairs from both models, alongside reference responses from the therapy transcripts dataset, as presented in Table 6.

In all three scenarios, the PFA-GPT model responses exhibit a clearer grasp of the emotional state and conversational context and can provide more specific, actionable, and empathetic advice compared to the GPT-3.5 responses. For instance, in response to a user message that complains about severe mood swings, the responses from the models are given below in Table 7:

An additional illustration is evident in the following scenario, shown in Table 8, where the models respond to ‘a user who complained about erectile dysfunction even after trying several medications and has a sexually frustrated girlfriend to whom he has suggested should have sex with other men.’

In the evaluation of PFA-GPT models, several criteria were assessed: relevance, personalisation, PFA (advice and

TABLE 4 Quantitative evaluation of GPT-3.5 conversational agent.

Scenario	BLEU score	ROUGE score			BERT score
		ROUGE-1	ROUGE-2	ROUGE-L	
1	0.0025	0.1930	0.0422	0.1769	0.8316
2	0.0222	0.2160	0.0310	0.1840	0.8380
3	0.0178	0.2364	0.0307	0.2091	0.8500
4	0.0293	0.2083	0.0285	0.1771	0.8450
5	0.0203	0.2822	0.0654	0.2183	0.8474
6	0.0140	0.2613	0.0426	0.2432	0.8554
7	0.0038	0.2267	0.0344	0.2024	0.8387
8	0.0096	0.2796	0.0228	0.2724	0.8397
9	0.0107	0.2183	0.0089	0.2013	0.8465
10	0.0002	0.1395	0.0173	0.1302	0.8315

Note: The values in bold are the best.

Abbreviations: BLEU, Bilingual Evaluation Understudy; BERT, Bidirectional Encoder Representations from Transformers.

TABLE 5 Quantitative evaluation of PFA-GPT conversational agent.

Scenario	BLEU score	ROUGE score			BERT score
		ROUGE-1	ROUGE-2	ROUGE-L	
1	0.0000	0.1930	0.0422	0.1769	0.8387
2	0.0163	0.2449	0.0371	0.2122	0.8437
3	0.0195	0.2986	0.0243	0.2805	0.8455
4	0.0197	0.2222	0.0424	0.2222	0.8486
5	0.0016	0.2323	0.0063	0.1717	0.8403
6	0.0337	0.2347	0.0360	0.2066	0.8482
7	0.0001	0.1622	0.0272	0.1622	0.8359
8	0.0039	0.2353	0.0401	0.2185	0.8404
9	0.0423	0.3060	0.0677	0.3060	0.8631
10	0.2227	0.2179	0.0775	0.2101	0.8360

Note: The values in bold are the best.

Abbreviation: BLEU, Bilingual Evaluation Understudy; BERT, Bidirectional Encoder Representations from Transformers; PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

TABLE 6 Comparison of the responses from the conversational agents to test user inputs.

S/N	Test user input	GPT-3.5	PFA-GPT
1	Teenager experiencing intense mood swings...	Addresses emotional volatility and coping.	Recommends keeping a mood journal.
2	Dealing with distressing voices...	Suggests seeking professional help.	Offers empathy and suggests professional support.
3	Coping with the pain of losing a baby...	Offers condolences and support suggestions.	Recommends seeking support groups.

Abbreviation: PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

guidance), professional referral, and validation and empathy, as shown in Table 9. PFA-GPT's responses excelled in terms of relevance, directly addressing user concerns with actionable advice and practical suggestions. Additionally, they exhibited a higher degree of personalisation, acknowledged the user's emotions, provided pragmatic guidance with specific steps, and frequently suggested professional assistance, all while displaying empathy and validation of the user's experiences.

Our study also incorporates a set of five ethical principles presented in Table 10, which are originally established by Coghlan et al. [47], including non-maleficence, beneficence, respect for autonomy, justice, and explicability, to guide the development of the mental health support system.

Table 10 shows that developed mental health support system focused on providing support while avoiding harm, respecting user autonomy, ensuring fairness, and promoting

TABLE 7 Comparison of responses from both conversational agents for scenario 1.

GPT-3.5	PFA-GPT
“It sounds like you're experiencing intense mood swings and a pattern of feeling high-spirited and motivated, but then becoming easily irritated and blaming others when things don't work out as planned. While I can't diagnose you, it's possible that you may be experiencing a combination of emotional volatility and difficulty managing frustration. This can be quite common during the teenage years due to hormonal changes and the challenges of navigating emotions and relationships. It's important to remember that everyone experiences emotions differently, and it's okay to feel a wide range of emotions. However, it's also important to find healthy ways to express and manage these emotions.”	“It sounds like your moods are fluctuating and you are noticing a pattern. You are aware of your moods and how you react to them. That is half the battle. I would suggest you start a mood journal. In the journal you will want to record your mood, the time of day, what you were doing, and who you were with. By having a record of your moods and the events leading up to it you can start to recognise your triggers. For example, if you notice that every time you are around a certain person it puts you in a bad mood then you know that is a trigger for you. Once you know your trigger you can figure out a plan to avoid it.”

Abbreviation: PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

TABLE 8 Comparison of responses from both conversational agents for scenario 2.

GPT-3.5	PFA-GPT
“I understand that you're facing challenges with erectile dysfunction, and it's commendable that you're seeking a solution to address this issue. However, suggesting an open relationship or giving permission for your partner to be intimate with other people is a decision that requires careful consideration and open communication between both partners. It's important to have an open and honest conversation with your girlfriend about your concerns and explore other ways to maintain intimacy and connection. Seeking guidance from a healthcare professional or a couple's therapist could also be beneficial in navigating this situation. Remember, every relationship is unique, and it's essential to prioritise open communication, empathy, and understanding to find a solution that works for both of you.”	“First of all, it is wonderful that you and your girlfriend are communicating with one another about what is going on in your lives. Open communication is crucial in maintaining a healthy relationship. However, I would advise against offering your girlfriend the option of having sex with other men. This could potentially complicate your relationship further with jealousy issues or trust issues. I would recommend asking your doctor about your erectile dysfunction. I am not a doctor, but there could be a simple explanation such as elevated blood pressure that you are not aware of. I would also recommend speaking with your doctor about alternative medications to treat your erectile dysfunction. There could be a medication that your doctor has not yet prescribed to you that may work.”

Abbreviation: PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

TABLE 9 Evaluation of conversational models based on domain-specific criteria.

Criteria aspect	GPT-3.5	PFA-GPT
Response relevance	Less directly relevant to user input.	More directly relevant, addressing specific issues.
Personalisation	Limited acknowledgement of the user's situation.	Higher level of personalisation, acknowledging emotions.
Advice and guidance	Offers general advice, lacks specifics.	Provides practical advice with actionable steps.
Professional recommendations	Rarely suggests seeking professional help.	Frequently recommends seeking expert assistance.
Validation and empathy	Limited empathy towards the user's feelings.	Includes more validation and empathy, fostering support.
Overall performance	Responses may lack depth and relevance.	Responses are tailored, empathetic, and actionable.

Abbreviation: PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

TABLE 10 Ethical appraisal of the conversational models.

AI-ethics principle	Ethical requirement	PFA-GPT
Non-maleficence	Prevent causing physical, social, or mental harm to users	Yes
Beneficence	Ensure interventions result in positive outcomes or genuine benefits for users	Yes
Respect for autonomy	Honour users' values and choices	Yes
Justice	Treat users impartially, avoiding bias, discrimination, or inequality	Yes
Explicability	Offer users adequate transparency about the technology's nature and effects, while being responsible for its design and implementation.	Yes

Abbreviation: PFA-GPT, Psychological First Aid-Generative Pre-Trained Transformer.

transparency. This approach aimed to offer PFA in a supportive and non-harmful manner while allowing users to engage voluntarily and providing multilingual support, all with

a clear understanding of the system's capabilities and limitations. It is important to note that this study was a proof of concept and did not involve handling actual user data or

interactions. To sum up, the findings substantiate the effectiveness and efficiency of the proposed mental health support system through the series of conducted experiments.

5 | DISCUSSIONS

The proposed mental health support system is a comprehensive platform designed to enhance emotional wellbeing and provide psychological assistance. The system comprises distinct components that work together to contribute to a holistic mental health support experience.

5.1 | Emotional distress detection

The emotional distress detection stands at the core of this proposed mental health support system as it is the backbone for the development of a chatbot that provides PFA. The system uses advanced deep learning techniques, particularly BERT, to accurately predict emotional distress levels in users' text inputs. The fine-tuned BERT model attained a remarkable accuracy of 93% and proved to be superior for the task of emotional distress detection compared to the Bi-LSTM which attained an accuracy of 78%. Several studies have used deep learning techniques like BERT for emotion detection using deep learning, the stacked CNN with LSTM architecture proposed by Khan et al. [57] achieved an accuracy of 71.67%, showcasing the diversity of human emotional responses. Similarly, Chiorrini et al. [58] attained 90% accuracy for emotion recognition in Twitter data using BERT, demonstrating the model's effectiveness in sentiment analysis. [59] proposed dual-stage architecture comprising BERT fine-tuning and Bi-LSTM classification for emotion detection using the ISEAR dataset. Their proposed model surpasses existing approaches, achieving 72.64% accuracy. Comparably, our fine-tuned BERT model achieved a promising 93% accuracy. Kumar and Jain [60] introduced the emotion-cause pair extraction technique to pre-process text data, achieving 98% accuracy with ECPE-BERT. While their approach enhanced BERT's performance, our study focused solely on emotion classification, attaining a strong accuracy of 93%. This emotional distress detection functionality of the proposed mental health support system enables the AI-powered chatbot to interact empathetically with users, offering PFA tailored to their emotional states. By analysing users' inputs and predicting distress levels, the chatbot provides appropriate responses, ensuring appropriate support and guidance during periods of emotional distress.

5.2 | Artificial Intelligence chatbot development for Psychological First Aid

The GPT3.5 model was made available for fine-tuning on the 22nd of August 2023 and this study implemented the fine-tuning using the therapy transcripts data, thereby creating a

novel fine-tuned model, PFA-GPT, for therapy-related tasks. The BLEU score and ROUGE score metrics have several limitations in the evaluation of generative conversational agents. This explains the extremely low scores computed using these metrics. Bilingual evaluation understudy score tends to prioritise exact word matches between generated and reference texts, overlooking the dynamic nature of human language where a single idea can be expressed in multiple ways. This approach can unfairly penalise responses that are accurate but not verbatim of the reference response. Bilingual evaluation understudy doesn't consider synonyms, paraphrasing, or changes in word order, which are common in natural language conversations. Recall-oriented understudy for gisting evaluation scores rely on n-gram overlap as their foundation, which means they do not capture the semantic nuances present in the text. Recall-oriented understudy for gisting evaluation overlooks factors like fluency, coherence, and contextual comprehension that are crucial for meaningful conversation. It places its primary emphasis on content overlap and disregards other vital dimensions. Recall-oriented understudy for gisting evaluation scores are also influenced by the length of the response, sometimes favouring longer outputs that coincidentally contain more matching n-grams. The BERT score was introduced to address the limitations of the BLEU and ROUGE scores. The BERT score captures contextually relevant responses, semantic and nuanced meanings beyond surface-level matching. Despite its advantages, the BERTScore metric undervalues fluency, coherency, and domain-specific dialogue intricacies, and may not fully capture the diversity of valid responses.

To overcome these limitations, a comprehensive evaluation strategy was proposed which involved a combination of these qualitative metrics, human judgements, and PFA-specific criteria. The comparison between the GPT-3.5 and the PFA-GPT model centred on their abilities to generate responses across diverse scenarios based on criteria such as relevance, personalisation, guidance, professional recommendations, validation, empathy, language structure, and overall user satisfaction. A standout discovery was the PFA-GPT model's exceptional proficiency in response relevance and customisation. Notably, it consistently offered PFA responses that were closely relevant to the test user inputs, surpassing the base GPT-3.5 model. The results produced by the PFA-GPT model also showed an increased level of personalisation and provided responses targeted to the user's specific circumstances and emotional state, fostering a deeper bond between the chatbot and the user. In terms of providing guidance and PFA, the responses provided by the PFA-GPT showed a proclivity for practical advice, coping mechanisms suggestion, active listening, and referral for professional care, particularly in scenarios involving interpersonal relationships, psychological wellness, and adaptive coping mechanisms. The fine-tuned model's responses revealed a better grasp of users' emotional states and concerns, translating into more actionable advice and suggestions. These chatbot responses frequently outlined the steps that the potential user could take to address their concerns in a practical manner. The models were evaluated in terms of their ability to suggest referral to mental health professionals. The

responses provided by the PFA-GPT model stood out for its consistent recommendation of obtaining advice from psychological professionals such as psychologists, therapists, and counselling services. This consideration is especially important in circumstances involving mental health, relational challenges, and grief. The responses provided by the PFA-GPT model show a clear expression of validation and empathy towards the user's feelings, sentiments, and experiences.

A compelling dimension of the comparison revolved around ethical considerations tied to the models' responses. Remarkably, the PFA-GPT model's responses aligned more cohesively with ethical criteria like non-maleficence, beneficence, respect for autonomy, justice, and explicability. The model's ability to generate relevant, empathetic, and actionable responses showcases responsible use, aligning effectively with established ethical guidelines for AI applications.

5.3 | Limitations and recommendations

Although the chatbot's responses produce promising outcomes, they are not a replacement for professional care. Additionally, performance could be affected by non-standard language, slang, and translation quality in the multilingual functionality's APIs. Another limitation is the absence of real-world user testing. This was not included in the current study because it falls beyond the study's scope. Involving participants to test the AI model is classified as a medical device software activity, which requires the involvement of clinical investigators, an ethical approval and an appropriate sponsor licence. Therefore, conducting real-time tests on patient participants falls outside the scope of the current phase of the project. These tests will be carried out in the next phase, once the necessary ethical approvals are obtained to allow participant interaction with the AI models.

Our current model also does not leverage Google Gemini, released by Google in 2024. Google Gemini offers enhanced capabilities, including improved natural language understanding, better contextual awareness, and the ability to process multimodal inputs that combine text, audio, and video data. This model represents a significant advancement over previous iterations and has the potential to handle complex and nuanced human interactions more effectively. Future research could explore integrating such advanced models for more nuanced human interactions, as they represent significant advancements over previous iterations.

Other future research opportunities include the following:

1. Exploring multimodal models that incorporate audio, video, and text data for enhanced emotional distress detection.
2. Expanding this proof of concept into a comprehensive web and mobile application with features like mental health diagnosis, PFA, virtual support groups, digital therapy, and professional referral recommendations.
3. Investigating advanced language models fine-tuned on extensive therapy data to cater to diverse mental health needs.

4. Conducting rigorous user testing to evaluate system performance and develop standardised qualitative analysis methods for assessing empathy and PFA in mental health chatbots.

5.4 | Potential limitations and unintended consequences of AI-powered chatbots in mental health support

While the integration of BERT-based emotional distress detection with the PFA-GPT chatbot demonstrates significant promise for AI-driven mental health support, it is essential to critically examine potential limitations and unintended consequences to ensure responsible and effective deployment in this sensitive domain.

• Risk of Over-Reliance on AI Systems

One key concern is the potential for users to develop an over-reliance on AI chatbots as their primary or sole source of mental health support. Although these systems can provide immediate assistance, they are not designed to replace professional therapists or psychologists. Users might bypass seeking professional help, assuming the chatbot is sufficient, which could delay necessary interventions for severe mental health conditions. To mitigate this risk, clear disclaimers and guidelines must accompany chatbot interactions, emphasising their role as a supplemental tool rather than a replacement for professional care.

• Challenges in Cultural and Contextual Sensitivity

Ensuring cultural and contextual sensitivity in chatbot responses is another critical challenge. Mental health experiences and expressions can vary significantly across cultures, languages, and individual contexts. Current AI systems, including the proposed PFA-GPT chatbot, rely on training data that may not adequately represent diverse populations. This limitation can lead to inappropriate or ineffective responses, further exacerbating feelings of misunderstanding or alienation in users. Incorporating diverse and representative training datasets, as well as mechanisms for continuous cultural adaptation, will be crucial in addressing this issue.

• Potential Biases in AI Training Data

Artificial intelligence systems are inherently dependent on the data used for their training. Biases in this data, whether due to underrepresentation of certain groups or historical inequities, can manifest in chatbot interactions. For example, emotional distress detection algorithms might misinterpret expressions of distress from marginalised communities due to insufficient representation in training data. These biases could lead to inequitable support outcomes, undermining the trustworthiness and ethicality of the system. To address this, rigorous audits of training datasets and regular evaluations of system outputs must be conducted to identify and rectify any biases.

• Importance of Human Oversight and Hybrid Models

Safety concerns are paramount when deploying AI chatbots in mental health settings. While these systems can provide valuable first-line support, they lack the ability to fully comprehend nuanced human emotions or handle complex psychological crises. Hybrid models that combine AI chatbots with human oversight can offer a balanced approach, ensuring that users with severe or escalating distress are promptly referred to human professionals. For instance, a tiered system where AI handles initial triaging and escalates high-risk cases to human intervention can mitigate risks while retaining the efficiency benefits of AI.

Incorporating insights from Dergaa et al. [5–7], we recognise the need to address nuanced attitudes towards AI deployment in sensitive domains. This work emphasises the importance of transparent communication with users, ethical safeguards, and ongoing monitoring to evaluate the chatbot's impact and effectiveness. Additionally, it highlights the role of user feedback in iteratively improving chatbot systems, ensuring they remain relevant, reliable, and respectful of user autonomy. By addressing these potential risks and areas for improvement, the integration of AI-powered tools like the PFA-GPT chatbot into mental health support can be guided by ethical and practical considerations. This underscores the importance of designing AI systems that prioritise inclusivity, transparency, and human oversight while maintaining their effectiveness and accessibility. These steps will not only enhance user trust but also ensure that AI-driven mental health solutions contribute positively and responsibly to this critical domain.

6 | CONCLUSION

This paper introduced a novel approach by integrating BERT and PFA-GPT to develop a mental health support system addressing these challenges, featuring emotional distress detection, a multilingual PFA chatbot, and a mood journal. The core of this mental health support system is its emotional distress detection system, powered by advanced deep learning techniques, particularly BERT, which achieved a remarkable 93% accuracy in predicting emotional distress levels in user text inputs compared to the Bi-LSTM model. This system allows the AI chatbot to interact empathetically with users, offering tailored PFA based on their emotional states, thereby ensuring appropriate support and guidance during periods of emotional distress. The developed PFA-GPT model outperformed GPT-3.5 in various aspects, offering more relevant, personalised, and practical responses that displayed empathy and validation. Notably, PFA-GPT's responses aligned more closely with ethical criteria, demonstrating responsible use and adherence to established ethical guidelines for AI applications. The proposed mental health support system offers promising results but has limitations, including potential variations in emotional distress expression across languages and cultures.

The study presents a promising AI-driven mental health support system, demonstrating significant advancements in

emotional distress detection and PFA delivery. The findings highlight the practical implications of integrating advanced AI models like BERT and GPT-3.5 in developing empathetic and ethical mental health chatbots. In real-world settings, this system can enhance accessibility to mental health support, providing immediate assistance and referrals to professional care when needed. However, as highlighted by Dergaa et al. [2, 3], it is crucial to acknowledge the current limitations of AI models in mental health applications and to continue refining these tools to ensure they meet clinical standards.

ACKNOWLEDGEMENT

For the purposes of open access, the authors have already granted a CC-BY licence over the author accepted manuscript to Keele University as per this Policy. This research did not receive any external funding.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data used in this study is publicly available, and the relevant references are provided in the dataset section.

ORCID

Baidaa Al-Bander  <https://orcid.org/0000-0002-2518-7364>

REFERENCES

1. Waisberg, E., et al.: A comparative analysis of ChatGPT and Google's AI's "bard" in medicine. *New Asian J. Med.*, 37–43 (2023). <https://doi.org/10.61838/kman.najm.1.2.5>
2. Dergaa, I., et al.: From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol. Sport* 40(2), 615–622 (2023). <https://doi.org/10.5114/biolSport.2023.125623>
3. Dergaa, I., et al.: Moving beyond the stigma: understanding and overcoming the resistance to the acceptance and adoption of artificial intelligence chatbots. *New Asian J. Med.*, 29–36 (2023). <https://doi.org/10.61838/kman.najm.1.2.4>
4. Washif, J., et al.: Artificial intelligence in sport: exploring the potential of using ChatGPT in resistance training prescription. *Biol. Sport* 41(2), 209–220 (2024). <https://doi.org/10.5114/biolSport.2024.132987>
5. Dergaa, I., et al.: ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Front. Psychiatr.* 14 (2024). <https://doi.org/10.3389/fpsyt.2023.1277756>
6. Dergaa, I., et al.: Evaluating the applicability and appropriateness of ChatGPT as a source for tailored nutrition advice: a multi-scenario study. *New Asian J. Med.* 2(1), 1–16 (2024). <https://doi.org/10.61838/kman.najm.2.1.1>
7. Dergaa, I., et al.: From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. *Front. Psychol.* 15, 1259845 (2024). <https://doi.org/10.3389/fpsyg.2024.1259845>
8. Koutsouleris, N., et al.: From promise to practice: towards the realisation of AI-informed mental health care. *Lancet Digital Health* 4(11), e829–e840 (2022). [https://doi.org/10.1016/s2589-7500\(22\)00153-4](https://doi.org/10.1016/s2589-7500(22)00153-4)
9. Coombs, N.C., et al.: Barriers to healthcare access among U.S. adults with mental health challenges: a population-based study. *SSM-Popul. Health* 15, 100847 (2021). <https://doi.org/10.1016/j.ssmph.2021.100847>
10. D'Alfonso, S.: AI in mental health. *Curr. Opin. Psychol.* 36, 112–117 (2020). <https://doi.org/10.1016/j.copsyc.2020.04.005>
11. World Health Organization: WHO: The World Health Report 2001: mental disorders affect one in four people. The World Health Report

- 2001 (2001). <https://www.who.int/news/item/28-09-2001-the-world-health-report-2001-mental-disorders-affect-one-in-four-people#:~:text=One%20in%20four%20people%20in,ill%2Dhealth%20and%20disability%20worldwide>
12. Czeisler, M.É., et al.: Mental health, substance use, and suicidal ideation during the COVID-19 pandemic — United States, June 24–30, 2020. *MMWR (Morb. Mortal. Wkly. Rep.)* 69(32), 1049–1057 (2020). <https://doi.org/10.15585/mmwr.mm6932a1>
13. Nobles, A.L., et al.: Identification of Imminent Suicide Risk Among Young Adults Using Text Messages (2018). <https://doi.org/10.1145/3173574.3173987>
14. Fitzpatrick, K., Darcy, A.M., Vierhile, M.: Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health* 4(2), e19 (2017). <https://doi.org/10.2196/mental.7785>
15. Inkster, B., Sarda, S., Subramanian, V.: An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth uHealth* 6(11), e12106 (2018). <https://doi.org/10.2196/12106>
16. Pruksachatkun, Y., Pendse, S.R., Sharma, A.: Moments of change. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13 (2019). <https://doi.org/10.1145/3290605.3300294>
17. Allen, S.V.: Artificial intelligence and the future of psychiatry. *IEEE Pulse* 11(3), 2–6 (2020). <https://doi.org/10.1109/mpuls.2020.2993657>
18. Faedda, G.L., et al.: Actigraph measures discriminate pediatric bipolar disorder from attention-deficit/hyperactivity disorder and typically developing controls. *JCPP (J. Child Psychol. Psychiatry)* 57(6), 706–716 (2016). <https://doi.org/10.1111/jcpp.12520>
19. Zheng, Y., et al.: Predicting beneficial effects of atomoxetine and citalopram on response inhibition in P arkinson's disease with clinical and neuroimaging measures. *Hum. Brain Mapp.* 37(3), 1026–1037 (2016). <https://doi.org/10.1002/hbm.23087>
20. Coppersmith, G., et al.: Natural language processing of social media as screening for suicide risk. *Biomed. Inf. Insights* 10, 117822261879286 (2018). <https://doi.org/10.1177/1178222618792860>
21. Hawa, S., et al.: Analysis of therapy transcripts using natural language processing. *Int. J. Eng. Adv. Technol.* 9(6), 489–494 (2020). <https://doi.org/10.35940/ijeat.f1598.089620>
22. Walsh, C.G., Ribeiro, J.D., Franklin, J.C.: Predicting risk of suicide attempts over time through machine learning. *Clin. Psychol. Sci.* 5(3), 457–469 (2017). <https://doi.org/10.1177/2167702617691560>
23. Vajawat, B., Varshney, P., Banerjee, D.: Digital gaming interventions in psychiatry: evidence, applications and challenges. *Psychiatry Res.-Neuroimag.* 295, 113585 (2021). <https://doi.org/10.1016/j.psychres.2020.113585>
24. Riches, S., et al.: Virtual reality relaxation for people with mental health conditions: a systematic review. *Soc. Psychiatr. Psychiatr. Epidemiol.* 58(7), 989–1007 (2023). <https://doi.org/10.1007/s00127-022-02417-5>
25. Park, M.J., et al.: A Literature Overview of virtual reality (VR) in treatment of psychiatric disorders: recent advances and limitations. *Front. Psychiatr.* 10 (2019). <https://doi.org/10.3389/fpsy.2019.00505>
26. Glaz, A.L., et al.: Machine learning and natural language processing in mental health: systematic review. *J. Med. Internet Res.* 23(5), e15708 (2021). <https://doi.org/10.2196/15708>
27. Eichstaedt, J.C., et al.: Facebook language predicts depression in medical records. *Proc. Natl. Acad. Sci. U.S.A.* 115(44), 11203–11208 (2018). <https://doi.org/10.1073/pnas.1802331115>
28. Corcoran, C., et al.: Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatr.* 17(1), 67–75 (2018). <https://doi.org/10.1002/wps.20491>
29. Bedi, G., et al.: Automated analysis of free speech predicts psychosis onset in high-risk youths. *Npj Schizophrenia* 1(1), 15030 (2015). <https://doi.org/10.1038/npjschz.2015.30>
30. Guntuku, S.C., et al.: Detecting depression and mental illness on social media: an integrative review. *Curr. Opin. Behav. Sci.* 18, 43–49 (2017). <https://doi.org/10.1016/j.cobeha.2017.07.005>
31. Mohapatra, S., et al.: Systematic stress detection in CNN Application. In: 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) (2022). <https://doi.org/10.1109/icrito56286.2022.9964761>
32. Kang, M., et al.: Classification of mental stress using CNN-LSTM algorithms with electrocardiogram signals. *J. Healthc. Eng.* 2021, 1–11 (2021). <https://doi.org/10.1155/2021/9951905>
33. Igbe, T., et al.: Deep learning intervention for health care challenges: some biomedical domain considerations. *JMIR mHealth uHealth* 7(8), e11966 (2019). <https://doi.org/10.2196/11966>
34. DiPietro, R., Hager, G.D.: Deep Learning: RNNs and LSTM, pp. 503–519. Elsevier eBooks (2020). <https://doi.org/10.1016/b978-0-12-816176-0.00026-0>
35. Bertl, M., Metsallik, J., Ross, P.: A systematic literature review of AI-based digital decision support systems for post-traumatic stress disorder. *Front. Psychiatr.* 13 (2022). <https://doi.org/10.3389/fpsy.2022.923613>
36. Uddin, M.Z., et al.: Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Comput. Appl.* 34(1), 721–744 (2021). <https://doi.org/10.1007/s00521-021-06426-4>
37. Siddik, S.A.N., Arifuzzaman, B.M., Kalam, A.: Psyche conversa - a deep learning based chatbot framework to detect mental health state. In: 2022 10th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia (2022). <https://doi.org/10.1109/icoict55009.2022.9914844>
38. Ji, S., et al.: MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare (2021). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2110.15621>
39. Brown, T.B., et al.: Language Models Are Few-Shot Learners (2020). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2005.14165>
40. Radford, A., et al.: In: Improving Language Understanding by Generative Pre-training. OpenAI (2021). Retrieved August 20, 2023, from https://www.researchgate.net/deref/https%3A%2F%2Fopenai.com%2Fblog%2Fchatgpt-plus?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIiwicGFnZSI6InB1YmxpY2F0aW9uRGV0YWlsIiwicHJldmVdXNqYXVldjIjYXRpb25EZXRhaWwifX0
41. Zhang, T., et al.: BERTScore: Evaluating Text Generation with BERT (2019). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.1904.09675>
42. Zhang, Y., et al.: DIALoGPT: Large-Scale Generative Pre-training for Conversational Response Generation (2019). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.1911.00536>
43. Ghanadian, H., Nejadgholi, I., Osman, H.A.: ChatGPT for Suicide Risk Assessment on Social Media: Quantitative Evaluation of Model Performance, Potentials and Limitations (2023). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2306.09390>
44. Lamichhane, B.: Evaluation of ChatGPT for NLP-Based Mental Health Applications (2023). arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2303.15727>
45. Zeberga, K., et al.: A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Comput. Intell. Neurosci.* 2022, 1–18 (2022). <https://doi.org/10.1155/2022/7893775>
46. Das, A., et al.: Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In: Proceedings of the BioNLP 2022 Workshop, Dublin, Ireland (2022). <https://doi.org/10.18653/v1/2022.bionlp-1.27>
47. Coghlan, S., et al.: To chat or bot to chat: ethical issues with using chatbots in mental health. *Digital Health* 9 (2023). <https://doi.org/10.1177/20552076231183542>
48. Rasheed, K., et al.: Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Comput. Biol. Med.* 149, 106043 (2022). <https://doi.org/10.1016/j.compbiomed.2022.106043>
49. Higgins, O., et al.: Artificial intelligence (AI) and machine learning (ML) based decision support systems in mental health: an integrative review. *Int. J. Ment. Health Nurs.* 32(4), 966–978 (2023). <https://doi.org/10.1111/inm.13114>
50. Potts, C., et al.: A multilingual digital mental health and well-being chatbot (ChatPal): pre-post multicenter intervention study. *J. Med. Internet Res.* 25, e43051 (2023). <https://doi.org/10.2196/43051>

51. Denecke, K., Vaaheesan, S., Arulnathan, A.: A mental health chatbot for regulating emotions (SERMO) - concept and usability test. *IEEE Trans. Emerg. Top. Comput.* 9(3), 1170–1182 (2021). <https://doi.org/10.1109/tetc.2020.2974478>
52. Rashkin, H., et al.: Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset (2018). *arXiv* (Cornell University). <https://doi.org/10.48550/arxiv.1811.00207>
53. Bertagnolli, N.: Counsel chat: bootstrapping high-quality therapy data. *Medium* (2020). <https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da>
54. Devlin, J., et al.: Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018)
55. OpenAI: GPT-3.5 (2023). <https://platform.openai.com/docs/models>
56. He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034 (2015)
57. Khan, A.N., et al.: Deep learning framework for subject-independent emotion detection using wireless signals. *PLoS One* 16(2), e0242946 (2021). <https://doi.org/10.1371/journal.pone.0242946>
58. Chiorrini, A., et al.: Emotion and Sentiment Analysis of Tweets Using BERT (2021). [Paper]. *EDBT/ICDT Workshops* <https://www.semanticscholar.org/paper/Emotion-and-sentiment-analysis-of-tweets-using-BERT-ChiorriniDiamantini/30159e346d89c1ce21275e2814e3c3f284991472>
59. Adoma, A.F., et al.: Recognizing emotions from texts using a bert-based approach. In: *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 62–66. *IEEE* (2020)
60. Kumar, A., Jain, A.K.: Emotion detection in psychological texts by fine-tuning BERT using emotion–cause pair extraction. *Int. J. Speech Technol.* 25(3), 727–743 (2022). <https://doi.org/10.1007/s10772-022-09982-9>

How to cite this article: Taiwo, O., Al-Bander, B.: Emotion-aware psychological first aid: Integrating BERT-based emotional distress detection with Psychological First Aid-Generative Pre-Trained Transformer chatbot for mental health support. *Cogn. Comput. Syst.* e12116 (2025). <https://doi.org/10.1049/ccs2.12116>