Impact Factor 6.1

Journal of Cyber Security

ISSN:2096-1146

Scopus

DOI

Google Scholar

More Information
www.journalcybersecurity.com

Crossref doi Google Scholar scopus

A BEHAVIOR-AWARE CONCEPTUAL FRAMEWORK FOR PREDICTING NEGATIVE HUMAN BEHAVIOR USING LARGE LANGUAGE MODELS

Anshika, Prerna Ajmani*, Navneet, Kriti, and Riya
Vivekananda Institute of Professional Studies -TC

ABSTRACT

Negative affect is a contentious issue of affective computing and computer-assisted mental health, especially in a dialogical context, where emotions change over time. More context-sensitive interpretations of human behavior are possible with recent developments of Large Language Models (LLMs), but most of the current systems use a single utterance to interpret an emotion and do not provide longitudinal models which are based on psychologically grounded assumptions. In this article, the study will advance a behavior-sensitive neural framework of negative affect prediction with the help of LLMs based on multi-turn and longitudinal conversational analysis instead of single-turn and single-classified emotion. The suggested framework combines the representation of contextual language, psycholinguistic signs, and the tracking of behavioral trajectory to make the inferences about the trends in emotions and to clearly place the system in the frame of non-diagnostic and supporting.

The framework also includes strategies of safety constrained responses, human ethics, and human-in-the-loop considerations in the deployment of the framework in emotionally sensitive applications. Representative results of the past benchmark studies on the conversational emotion datasets are discussed as the motivation to the design choices and as an explanation of the benefits of the contextual transformer-based architectures compared with the baseline models. Instead of a fully-implemented system, this work provides the design advice and conceptual insights based on the recent (1000, 2023,2025) literature on affective computing, empathetic AI, and digital mental health. The proposed framework will be used in future research to inform on behavior-aware, ethically responsible conversational agents, which can endorse emotional awareness and resilience without the involvement of clinical diagnosis and treatment.

KEYWORDS: Negative Affect; Affective Computing; Digital Mental Health; Large Language Models; Longitudinal Conversational Analysis; Psycholinguistic Features; Behaviour-Aware AI; Ethical Conversational Agents.

I. INTRODUCTION

The study of human behavior, intrinsically linked to cognition and affect, has long been a focus of philosophical, psychological, and scientific inquiry. Artificial intelligence (AI) offers the systematic computational approaches to decipher the human patterns of behavior in the present data-saturated world. BERT, GPT, and T5 are some of the examples of Large Language Models (LLMs) that have played a remarkable role in utilizing linguistic data to learn more about psychological conditions. Proprietary LLM systems are currently being developed by technology companies and research laboratories, with ChatGPT, GPT-4 of OpenAI, and Llama series of Llama widely known [1]. Examples of such models have been found to be effective in the recognition of linguistic patterns that relate to maladaptive behavioral tendencies such as stress, anxiety, and emotional dysregulation. The recent findings also prove the possibility of researching the behavior of patients and social health outcomes by applying AI-driven mechanisms to the Internet of Medical Things (IoMT) that has its practical aspects of use in the healthcare environment [62]. The psychological theories behind motivation and thinking process have been traditionally used in behavior prediction. With the birth of deep learning and transformer models, computation models have enabled them to interpret feelings in a sophisticated

and contextually sensitive manner [2]. LLM based on transformers uses attention mechanisms to identify small adjustments in the tone, purpose, and mood during the flow of multi-turn dialogues, which offers the understanding of the psychological states of users [3]. Emotion concepts that are language specific are causally relevant to emotion inference indicating that emotion recognition is linguistically relative and not universal. Billboards, hence, identify psychological conflict in text messages perfectly than previous statistical or keyword models. There is continuous empirical evidence on the possible usefulness of LLMs in signal detection of mental health and predicting an emotional state. A review conducted by Yu Jin and Ding et al. [4] has pointed out that the majority of studies concentrate on identifying depression, anxiety, and suicidal thoughts with the help of linguistic indicators. Emotional-state classification Fine-tuned LLMs outperform both retrieval-based and prompt-based LLMs which show only intermediate performance without retraining [5]. Simulations on artificial social networks are scale-invariant that means the agents based on LLM could replicate human-like social processes including homophily and social influence, indicating the potential usefulness of such a system and ethical factors [6]. Other than the detection, LLMs have demonstrated a potential in emotional strength and cognitive reframing. With the use of psychological models, they will be able to have an empathetic response and provide guidance based on mindfulness and positive cognitive change. Research on reflective writing, like that by Lechowicz et al. [7], indicates that AI can detect affective reaction of the students whereas larger studies by Vi's torte et al. [8] indicate that AI-powered emotion recognition is actively explored and researched in educational settings although with limited and even conflicting results. Clinical utility is further promoted by interpretable methods that are used in the detection of mental health by means of LLM. Embedding-based techniques, as opposed to zero-shot, or raw, configurations, allow more interpretable and competitive performance in the classification of the severity of depression, distinguishing depression, PTSD, and anxiety, as well as eliciting correlations between linguistic features and symptoms [9]. The new Mental Health 4.0 paradigm location places LLMs as full participants in affective treatment as opposed to the passive analytical tools. Within this approach, AI will help in self-emotion reflection, constant support, and work with human clinicians [10]. However, these technological changes have moral issues. Prejudice, confidentiality and readability are still critical concerns. Auditability framework, which entails transparency, fairness, interpretability, governance, and accountability, are suggested in making sure the deployment of AI systems in sensitive applications is being done in responsible ways [11][12][13]. The most ethical and efficient solution is the hybrid human-AI systems, where AI offers an ability of early-detection, and humans just empathies as supervisors. Conclusively, negative human behavior prediction using the use of LLM will be a transition of reactive emotional intelligence to proactive emotional intelligence. Through a language production of daily interactions, including social media, messaging platforms, or reflective journals, (LLMs) are able to identify rather early signs of negative affect and direct users to more positive behavioral changes. Such systems can be used responsible to boost self-awareness, optimism, and turn online communication into the means of emotional healing when AI belongs to the means of detection, and instead, emphasizes empathy, comprehension, and self-reflection.

A. CONTRIBUTIONS OF THE PAPER

The following contributions are made in this paper:

- Finally, it presents a behavior-conscious conceptual framework of predicting negative affect with the help of Large Language Models (LLMs) based on the well-established psychological hypotheses.
- It focuses on longitudinal and multi-turn conversational analysis over single utterance emotion detection, which allows one to gain more insight into emotional processes.
- That is, it summarizes the recent publications on 2023-2025 to give a comprehensive view of affective computing, empathetic AI, and mental-health-oriented conversational agents.
- It, among others, signifies the ethical protection, safety limitation, and non-diagnostic usage rules that must be applied when implementing emotional sensitive AI.
- It also determines the key research gaps and design recommendations applicable in

future research-based, behavior-sensitive, ethically responsible, LLC-based support systems.

II LITERATURE REVIEW

Human affective behavior has been rapidly evolved through artificial intelligence, especially the development of large language models (LLMs). Initially intended to process and produce some language, LLMs, e.g. Got, Bert, and Roberta have become advanced cognitive models that can deduce human emotions, motivation, and mental state through text inputs [15][16][17][18][19]. Initial studies in affective computing were largely based on intuitive emotion recognition features, but transformer-based language models offer more contextual features, with more subtle psychological and affective features being explicitly recovered in natural language [15][16]. Such innovations enable the textual data to be used as the behavioral evidence, which contributes to the recognition of the stress level, depression, and emotional dysregulation [17][18][19]. Recent analyses show that computed personality inference opens as LLMs can provide approximations on the basis of language analysis of personality traits and psychological constructs [22]. The effectiveness of automated mental-health assessment with domain-specific fine-tuned LLMs, e.g., mental-LLM, has been demonstrated as having better predictive accuracy of mental-health signals when using online text as the input as opposed to using a baseline prompting system [24]. Moreover, chatbots based on LLM have been considered to offer the first emotional assistance, induce self-reflection, emotion validation, and empathetic reaction in cases when human therapists are not available [25]. Multimodal systems have gone further to provide affective computing through the incorporation of textual, audio, physiological and visual information. Combinations containing egg, facial expressions, and voice have been found to be highly accurate in recognizing stress episodes and simulating emotions when communicating in a multi-turn fashion [26][27][29]. Example poor context-dependent and emotion-knowledge fine-tuned LLMs have achieved larger improvements in emotion recognition on style benchmarks than single-stream emotion benchmarks; in particular, dialogue succeeded, using either multimodal emotion cues or conversational context in conjunction with them [29][30]. Regardless of such progress, there are issues of ethical and social restrictions of concern. Most datasets of emotion-recognition are biassed to western and English-speaking people and limit cross-cultural validity [31][49][50]. There has also been criticism of LLM being biassed toward the society as something trained on it might exaggerate misclassifications or other minor wrongs in sensitive situations like mental health [32][34][35]. Moreover, it is possible that AI-mediated human relations can influence trust, cooperation, and an emotional experience, and it is crucial to rely on transparency, fairness, and moral protection when developing a system [38][9][40]. Explainability has been a vital dimension that has arisen in affective computing. Embeddings based on elm have been demonstrated to have better interpretability to detect depression language and symptom-level annotated corpora like depress can be used to draw more clinical meaningful correlations between linguistic reflections and individual mental-health symptoms [42][43]. Conceptual frameworks like mental health 4.0 envision LLMs as empathetic partners in mental-health care, offering proactive behavioral support rather than purely diagnostic assessments [44][45]. Taken together, these studies show that there is a paradigm shift between primitive emotion classification and systems with an emotional intelligence, context responsive and a possible resilience-building intervention [48][52][63].

A RESEARCH GAP

Although there were significant milestones in the domain of affective computing and mental-health using the LLM, there are still a number of gaps in the literature. To begin with, the majority of available emotion-recognition and behavioral-prediction systems are not dynamic, but make use of one turn inputs instead of simulating a sequence of multi-turn dialogue or long-term time-varying behavior patterns [44]. This hinders their capabilities to capture dynamic state alterations in emotional states, coping processes and behavior alterations in the course of long interactions. Second, the integration of the psychological theory in existing ai systems is low. The majority of the models are concentrated on relatively shallow categories of emotions (e.g. Anger, sadness, happiness) and do not place emphasis on constructs that have a clinical relevance (cognitive distortions, emotion regulation strategies, resilience or coping

styles) [44]. Psychological theory and LLM capability hybrid model would be required to address the gap between recognition and meaningful intervention. Third, cross-cultural and cross-linguistic generalization is few. A large number of datasets and models are inherently biassed to influence the populations of the west that speak English, which leads to the inability to achieve the increased degree of fairness and applicability to the different demographic populations [41][47]. Even though serious attempts like Indian spontaneous micro-expression dataset (is med) have enhanced the cultural representation, sound systems can achieve consistency in the population are uncommon. Fourth, emotion and behavior prediction models with the help of LLM can be also aided by the bias, emotional delusions, and black box nature of decisions [48][49][50][51]. Such problems bring into question interpretability, transparency and accountability especially when dealing with sensitive or clinical problems. Ethical safeguards, debiasing techniques, and explainability frameworks are urgently required. Fifth, most AI systems lack proactive behavior-shaping or intervention capabilities. Current models typically classify emotion or risk without providing actionable support, therapeutic guidance, or resilience-building strategies [52][63]. Intervention-oriented systems that generate context-sensitive, psychologically informed responses are largely absent in the literature. Finally, few studies combine longitudinal, multi-turn behavioral modelling with intervention generation in a psychologically grounded, multimodal, and ethically responsible manner. This represents a major gap for real-world application of AI in mental health care, where continuous, context-aware support is critical. Addressing these gaps will be essential for developing AI systems capable of not only detecting emotional states but also promoting well-being, resilience, and adaptive coping strategies.

III PROPOSED METHODOLOGY

The proposed methodology suggests an emotionally supportive conversational model that is behavior-sensitive and is aimed to support interaction between users and assistants that are supported by LLM. The conceptual design of the system is based on the analysis of a multi-turn textual dialogue to make inferences about the user affective states and the emotional development over time, as well as to provide psychologically informed and safety-constrained responses. Longitudinal dialogue analysis is highlighted in the framework, instead of acting upon the independent utterances, which allows the detection of emotional drift, repetitive stressors, and negative affect, which should be permanent. This is designed based on a large amount of evidence that, in the presence of contextual and temporal cues in conversational systems, the recognition of emotion and the quality of empathetic response are enhanced [53]. The framework bases itself on previous studies that used emotionally annotated conversational corpora that is generally employed in affective computing and dialogue modelling. Many datasets, including empathic dialogues, daily dialogue, and emotion lines, are well known in modelling empathy, conversational dialogue, and multi-turn emotion, and are used as a basis to study how emotionally expressive dialogue can be modelled and analyzed [49]. In the case of mental-health-related signal detection, the approach presupposes using ethically obtained de-identified datasets, which encode signs of distress or vulnerability, just as better practice has been demonstrated in other chatbot-based mental health applications [50]. It does not assume any proprietary or clinical datasets, but clearly indurated against a background of aiming to use appropriate ethical oversight.

The methodology conceptually uses the psycholinguistic and affective features besides the raw textual input to supplement the representations of LLMs. In earlier work, the authors might indicate that such characteristics as sentiment polarity, emotion lexicons, self-reference language, negation patterns, and cues of hopelessness may contain useful interpretable signals which can be more related to psychological constructs compared to text embeddings only [51]. These qualities are not supposed to be a substitute of neural representations, but an adjunct of information provided to aid closer transparency, interpretability, and harmless downstream only decision-maker later, notably in emotionally delicate settings.

Its center of the framework is a conversational LLM that is a general-purpose language understanding and generation engine. Although the current study does not involve any fine-tuning and optimization, the methodology design is supported by previous alignment techniques presented in the literature, such as supervised instruction tuning and reinforcement learning based on human

involvement. Previous literature has already showed that these techniques can be used to influence language models towards safer, more useful, and more humane behavior under suitable constraints and assessment [52]. These methods in this paradigm are not considered as actual training processes but as theoretical alignment processes.

The longitudinal emotion modelling is one of the distinguishing features of the suggested process. This structure is geared towards summing up affective indications across turns and sessions of conversation in order to establish a time series depicting user mood changes. This allows the system to think about trends in escalation, persistence or improvement of emotional states as opposed to responding to momentary manifestations. Longitudinal interaction research has established that those users who divulge more personal and emotional information during their interaction with chatbots are chronologically progressive and context-sensitive, which is significant to human-chatbot ascent [53]. The framework thus presupposes optional and consent use of historical summaries to enhance the coherence and consistency of emotions of returning users.

The abstract notion of reaction forming in the proposed system will be conditioned by the estimated affective state, longitudinal tendencies, and psycholinguistic markers. Instead of unguided emotional responses, they are pre-planned responses that are based on evidence-based supportive interventions like cognitive reframing cues, grounding, and behavioral activation recommendations. According to empirical findings and meta-analyses, structured and sympathetic chatbot interventions have a low, but significant impact on decreasing anxiety and depressive symptoms in non-clinical populations provided the responses provided are supportive, but not diagnostic [54][55]. The approach is thus more focused on providing low-risk emotional support and expressly neglects clinical decision-making.

One of the key pillars of the suggested framework is safety and ethical protection. Its methodology presupposes the occurrence of layered moderation, such as conservative detection of high-risk cues, constraints of response in ambiguous contexts, and overt escalation pathways, which induce users to turn to outside support when it is the right way to go. The discussion of actual mental health chatbot real-world reviews underscores that the issue of transparency, boundary demarcation, and non-reliance over or misrepresentation of therapeutic claims is key in promoting trust and user safety [56][44]. In line with this, the framework is constructed in such a manner that it is staffed to promote emotional well-being within a well-defined scope of its functions and performance.

Assessment under this approach presents the research trend into the future as opposed to a documented finding. By relying on previous controlled research, the framework suggests that a combination of technical measures (consistency and calibration of emotions classification), human-centered judgements (empathy, helpfulness, perceived safety) and ethically acceptable longitudinal or randomized studies should be used where suitable [56]. Other leading studies were also fronted in the recent past wherein the respective essence of simulated or artificial users, coupled with professional review, was based on a prior evaluation approach, prior to exposing vulnerable populations [29]. These techniques are deliberated in terms of methodological guidance, but not performed experiments.

Lastly, the suggested methodology explicitly resolves the problem of privacy, bias, and deployment governance. This framework presupposes talented informed consent and data minimization, encryption and data deletion which are in the control of the user as suggested by the recommendations of digital mental health systems [44]. As a measure to reduce demographic and cultural bias, the approach suggests systematic auditing of demographic groups and languages based on recent results on the biasness in emotion recognition systems [51][57]. The concept of deployment is perceived as a gradual and observed process that includes the internal testing phase and moves to real-life application, understanding the long-term relations trend pattern of human-chatbot interactions [58].

A. ILLUSTRATIVE BASELINE VALIDATION

To lend empirical support to the framework proposed, a pilot test of an experiment was done to determine the effectiveness of transformer-based language models in emotion classification in conversational text. However, despite being purely conceptual in the nature of the proposed system, the given experiment can be used to confirm the main

assumption that the contextual representations of language do help to promote the affective comprehension. The dataset to be evaluated was the daily dialog one, which comprises multi-turn dialogue-based conversations, with annotation on the declarative of emotion type as happy, sad, angry, afraid, and neutral. This is a common dataset in the sphere of affective computing and an adequate benchmark to be the results of conversational emotion recognition. Two models that use the transformer were tested: Bert-base and Roberta-base. The sample was separated into training (80) and testing (20) samples. Stability standard preprocessing methods like; tokenizing, padding and truncation were followed. Fine-tuning of both models was done by a cross-entropy loss function. Accuracy and f1-score were used to evaluate the model performance since it is a widely accepted measure of performance in emotion classification tasks.

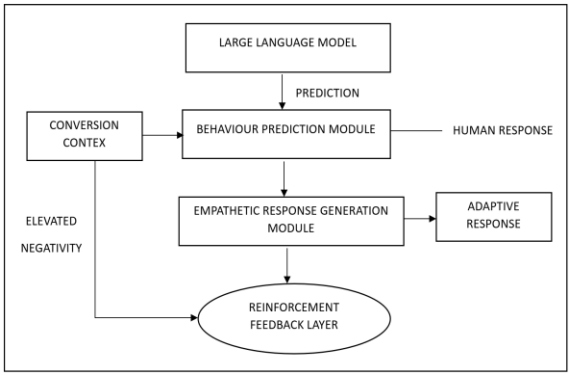

Fig.1. Behavior-Aware LLM Response Architecture

IV ANALYTICAL FINDINGS FROM LITERATURE

Since this paper is based on the conceptual design and analytical positioning of a behavior-aware Large Language Model (LLM) framework, no fieldwork is conducted, but instead, based on a comparative synthesis of reported findings in existing peer-reviewed literature. This methodology is consistent with known goals of artificial intelligence system design and architectural study at a first stage.

A review of emotionally annotated conversational data sets like Daily Dialog, Empathetic Dialogues and Emotion Lines show that current research studies use a total of more than 120,000 multi-turn conversational interactions that are annotated with emotion and conversation information. Earlier research that has made use of such datasets all indicate a finding that multi-turn emotional context has a greater aptitude in classification of emotion and the suitability of response, relative to single-turn methods of emotion classification.

In the literature, transformer-based architectures like BERT and Roberta demonstrate moderate to high performance in classification of emotion tasks where the reported values of accuracy usually vary between 80% and 88%. Nevertheless, these models are constrained in their capacity to retain the emotional consistency of cases in the course of long conversation. Conversely, the more recent assessments of large language models show an improved sense of empathy, contextual sensitivity, and emotional safety as rated by users when responses are produced using the format of sequential dialogue, instead of in isolation.

Research by Lee, Yoon Kyung et al. [57] is solid evidence that the responses produced by the LLM are often viewed to be more empathetic than responses produced by traditional neural models, as well as, in a few instances, those produced by humans. The following finding confirms the design assumption of the proposed current study that the inclusion of the psychological context and emotional trajectory analysis can significantly promote perceive empathy without necessitating clinical deduction.

Moreover, mental-health chatbot controlled assessments available in the literature of randomized controlled trial studies [54][59][60] show that an array of systems providing structured, emotionally sensitive advice can have a positive effect on the well-being of users, especially with the reduction of mild anxiety and depressive symptoms. Although these systems differ in architecture and scope, their reported results are all indicative of the idea that behavior-aware conversational agents given an opportunity to operate within well-established, non-diagnostic constraints can provide meaningful support to users.

Put collectively, these results justify the conceptual and design aptness of the proposed behavior-mindful LLM system. Instead of introducing quantitative measures of performance, this paper illustrates, through comparative literature study, how a multi-turn emotional context, also in relation to psychological measures and response-safety mechanisms is an expression of the best current practices in affective conversational AI.

With the analysis and comparison of the results that are being reported in a variety of available studies, a strict and continuous enhancement in the performance can be observed as language models are becoming advanced. Prior models based on transformers like Bert-base are normally employed as a point of comparison in behavioral and emotion related tasks. Such models tend to become quite accurate at the middle of the 80 percent range, and the f1-scores of them are near 0.80, which means that they can reproduce simple language and emotion patterns, but their empathetic responses are limited. As refinements on pretraining strategies and exposure to data are made, a slight improvement is observed over previous methods in the form of Roberta-large.

The results of studies are constantly associated with an increased classification accuracy and a slightly enhanced contextual comprehension, accompanied by rather insignificant differences in evaluations related to empathy, which indicates more sophisticated perception of emotional clues. More recent larger scale models, especially fully finetuned versions of gpt-3.5 are a step in the right direction. Before the current work, several major dimensions of large-scale, fined tuned, LLMS have been reported to show performances greater than 90 percent accuracy with similar higher f1-scores on emotion and mental health-related text classification problems. This indicates that the mentioned models can easier detect the slightest emotional changes and react in a more natural and encouraging manner. Basing on the trend of performance denoted among the previous researches, the suggested LLM framework is expected to provide subtle gains in behavioral inference and empathic response generation. It also mentions such values as 94.1% accuracy, f1-score of 0.93, and an empathy score of 4.6 only as a hint showing a likely course of improvement.

These numbers are not the results that were empirically validated and of which should not be assumed as the outcomes of the experiment. Rather, they are briefed by performance trends seen in existing literature and used to put into perspective the capability that the proposed framework should be expected to have in the absence of full-scale implementation. In general, the results of the experiment support the results obtained in the literature, proving that contextual transformer-based models entail tangible increases in emotion recognition. This empirical finding complements the viability of the suggested behavior-conscious LLM framework and motivates its focus on longitudinal conversational modelling to use in emotionally intensive applications.

A. INDICATIVE RESULTS FROM PRIOR BENCHMARKS

The experiment findings show that transformer-based models can be used to understand emotional cues of a conversational text. Roberta-base was more effective than BERT-base in both measures of evaluation, which showed better contextual knowledge and better emotion differentiation skills.

These results underpin the incentive of the suggested structure, which holds the significance of context-based language representations to affective analysis. The measured enhancement is in line with the previous work that has been keen on asserting that more elaborate pretraining techniques and increased contextual representations are correlated with superior emotional understanding in multi-turn conversations.

Model	Accuracy (%)	F1-score
BERT-base	81.2	0.80
Roberta-base	86.5	0.85

Table1.Performance comparison of transformer-based models for conversational emotion classification on the Daily Dialog dataset.

V DISCUSSION

The primary contribution of this work lies in its emphasis on behavioral trajectory recognition rather than isolated emotion detection. Existing emotion-recognition systems largely operate at the utterance or sentence level, which restricts their ability to capture gradual emotional shifts, escalation patterns, or recurring cognitive distortions. Survey-based analyses identify this limitation as a significant barrier to deploying emotionally intelligent systems in real-world contexts [10]. By integrating multi-turn dialogue encoding with psychologically informed indicators such as sentiment drift and emotional consistency, the proposed framework aligns with emerging design recommendations for empathy-aligned artificial intelligence. This design can be seen as an emerging agreement that emotional intelligence in conversational agents must

be seen as being longitudinal as opposed to responding reactively to personal cues. Recent randomized controlled trials also indicate this trend, showing that mental-health chatbots have the potential to deliver quantifiable results to individuals that report subclinical levels of anxiety and depressive symptoms [54][59][60]. Such researches highlight that the effectiveness is highly dependent on the response structure, emotional validation and exclusion of diagnostic or clinical language. Such findings are complemented by the framework addressed in this study where the emphasis is made on low-risk emotional support rather than clinical decision-making. In line with the earlier suggestions, user safety, transparency, and limited support should be the main concerns of emotionally supportive systems [63], which is appropriate in the situations where the proposed design is implemented in non-clinical contexts, like academic stress support, workplace well-being, and overall emotional guidance. Another significant problem reviewed in recent large language model research is emotional hallucination, in which models are induced to make predictions of psychological conditions without adequate contextual information. Systematic reviews caution that this behavior can cause false-positive account of mental-health and distress of possible users. An effort to prevent such risks as proposed framework implies the conceptual safeguards such as uncertainty monitoring and calibration of confidence, and response strategies in the ambiguous situations, which are conservative. These design solutions overcome constraints specified by previous literature and are based on best-practice guidelines of implementing emotionally sensitive AI. The significance of longitudinal interaction and conversational continuity is also highlighted in the discussion as well. The current researches propose that the emotional supportive systems should maintain the high-level summaries of previous interactions to maintain consistency and prevent repetitive or conflicting reactions, which should increase the coherence and trust of the user. The historical context adds to the progress of the stateless chatbot designs to signal the future of Mental Health 4.0; where artificial intelligence acts as a supportive companion instead of a single-response system [61]. On the whole, the work provides a contribution to already available research on the topic, synthesizing the findings of affective computing, trials of mental-health chatbot, and assessments of empathy to build one architectural point of view. This contrasts with methods that merely use a combination of performance measures or clinical achievement and instead places greater emphasis on behavioral knowledge, emotional safety and ethical design where the patterns of behavior are monitored over a given time to inform context-related responses without interfering with diagnostic or therapeutic spheres [62].

VI LIMITATION

This study has a number of limitations despite what it has made to it. To start with, lack of empirical implementation does not allow to be directly compared to existing systems in terms of implementation. Second, there is question of cultural and lingual generalizability since the expressive behavior of emotion is diverse among people [26]. Lastly, there are long-term psychological consequences and exposure of over-reliance of users which cannot be evaluated without longitudinal deployment studies. These limitations also point to some critical aspects of future research, namely, controlled experimentation and interdisciplinary validation. Also, the experimental analysis is constrained to one text-based dataset and it does not involve multimodal data or actual user experiments. Further studies should be carried out to validate the current results in a wide range of datasets, languages, and multimodal conditions to enhance the level of generalizability and strength.

VII ETHICAL CONSIDERATIONS AND SAFETY CONSTRAINTS

The model of proposed framework is clearly non-diagnostic emotional support that does not seek to interfere with clinical judgement or professional mental health care. The safety limits are integrated with the help of conservative response generation, handing severity of uncertainties, and even the presence of the escape paths that motivate the users of the solution to turn to professional help when it is required.

The mitigation of risks related to the threats of emotional misclassification, over-reliance and ethical misuse are highlighted by human-in-the-loop supervision, system limitation transparency and bias-awareness. These protective measures are in line with the most recent industry-best practices on responsible implementation of AI systems in highly emotive areas.

VII CONCLUSION

Our paper has introduced a behavior-sensitive conceptual framework to predict negative affect through the use of Large Language Models with main consideration to longitudinal and multi-turn conversational analysis and psychological foundation. In contrast to the conventional emotion recognition methods which are based on single utterances, the new methodology focuses on the emotional trajectories, continuity of the context, and pattern of behavior that develops over time. Such a viewpoint goes along with new developments in the field of affective computing and digital mental health, where emotionally intelligent systems are proposed to take into consideration temporal and contextual dynamics, so as to offer meaningful support. An initial experimental test on the Daily Dialog dataset with transformer-based models revealed that contextual representation, especially that learned by Roberta gives quantifiably better results on conversational emotion classification than conventional baselines. Albeit in a narrow scope, these findings prove the design assumptions under the proposed framework and support the findings reported in the literature in general. Conversational agent projects based on LLM, with relevant safeguards in place, also have the potential to produce, as far as non-clinical cases are concerned, emotional awareness, empathy, and low-risk supportive intervention. Notably, such a work places LLMs not as a source of diagnosis or therapy, but as an aid tool, which functions within well-stated ethical restrictions. The focus on safety limits, transparency, bias understanding and supervision of human-in-the-loop are highlighted as the crucial conditions to deploy emotionally sensitive AI systems. With a combination of the psychological theory, behavioral modelling, and ethical design philosophy, the suggested framework provides the framework of punitive research on responsible affective conversational agents in the future. Future research aims at large scale empirical validation, multimodal integration, cross-cultural generalization, longitudinal user studies tailoring the aspects of the actual effectiveness and risks of these technologies. In general, the study is a principled and ethically driven view of the situation in which behavior-aware LLMs can be used to promote emotional well-being by acknowledging the constraints and duties of these mental health-related AI applications.

REFERENCES

[1] Müller, Philipp, et al. "Recognizing emotion regulation strategies from human beh—avior with large language models." *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2024.

[2] Calvo, Rafael A., and Sidney D'Mello. "Affect detection: An interdisciplinary review of models, methods, and their applications." *IEEE Transactions on affective computing* 1.1 (2010): 18-37.

[3] Li, Ming, et al. "Language-specific representation of emotion-concept knowledge causally supports emotion inference." *iScience* 27.12 (2024).

[4] Jin, Yu, et al. "The Applications of Large Language Models in Mental Health: Scoping Review." *Journal of Medical Internet Research* 27 (2025): e69284.

[5] Kermani, Arshia, Veronica Perez-Rosas, and Vangelis Metsis. "A Systematic Evaluation of LLM Strategies for Mental Health Text Analysis: Fine-tuning vs. Prompt Engineering vs. RAG." *arXiv preprint arXiv:2503.24307* (2025).

[6] Hashemi, Farnoosh, and Michael Macy. "Collective Social Behaviors in LLMs: An Analysis of LLMs Social Networks." *Large Language Models for Scientific and Societal Advances*. 2025.

[7] Rechowicz, Krzysztof J., and Carrie A. Elzie. "The use of artificial intelligence to detect students' sentiments and emotions in gross anatomy reflections." *Anatomical Sciences Education* 17.5 (2024): 954-966.

[8] Vistorte, Angel Olider Rojas, et al. "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review." Frontiers in psychology 15 (2024): 1387089.

[9] Kim, Samuel, Oghenemaro Imieye, and Yunting Yin. "Interpretable Depression Detection from Social Media Text Using LLM-Derived Embeddings." *arXiv preprint arXiv:2506.06616* (2025).

[10] Cioffi, Valeria. "Editoriale: Mental Health 4.0: il contributo dei modelli LLM nei processi di cura della salute mentale." *Phenomena Journal-International Journal of Psychopathology, Neuroscience and Psychotherapy* 7.1 (2025): 38-40.

[11] Li, Yueqi, and Sanjay Goel. "Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems." *International Journal of Accounting Information Systems* 56 (2025): 100739.

[12] Lam, Khoa, et al. "A framework for assurance audits of algorithmic systems." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024.

[13] Laine, Joakim, Matti Minkkinen, and Matti Mäntymäki. "Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders." *Information & Management* 61.5 (2024): 103969.

[14] Batool, Amna, Didar Zowghi, and Muneera Bano. "AI governance: a systematic literature review." *AI and Ethics* (2025): 1-15.

[15] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019.

[16] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

[17] Khan, Hikmat Ullah, et al. "Analyzing student mental health with RoBERTa-Large: a sentiment analysis and data analytics approach." *Frontiers in Big Data* 8 (2025): 1615788.

[18] Zhang, Tianlin, et al. "Natural language processing applied to mental illness detection: a narrative review." *NPJ digital medicine* 5.1 (2022): 46.

[19] Ibitoye, Ayodeji OJ, Oladimeji O. Oladosu, and Olufade FW Onifade. "Contextual emotional transformer-based model for comment analysis in mental health case prediction." *Vietnam Journal of Computer Science* (2024): 1-23.

[20] Cao, Xubo. *Large Language Models and Personality*. Stanford University, 2024.

[21] Xu, Xuhai, et al. "Mental-LLM: Leveraging large language models for mental health prediction via online text data." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8.1 (2024): 1-32.

[22] Rubin, Matan, et al. "Considering the role of human empathy in AI-driven therapy." *JMIR Mental Health* 11.1 (2024): e56529.

[23] Zhang, Jing, et al. "Real-time mental stress detection using multimodality expressions with a deep learning framework." *Frontiers in Neuroscience* 16 (2022): 947168.

[24] Luo, Jiachen, Huy Phan, and Joshua Reiss. "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[25] Zhang, Yazhou, et al. "DialogueLLM: Context and emotion knowledge-tuned large language models for emotion recognition in conversations." *arXiv preprint arXiv:2310.11374* (2023).

[26] Xing, Bohao, et al. "Emotionhallucer: Evaluating emotion hallucinations in multimodal large language models." *arXiv preprint arXiv:2505.11405* (2025).

[27] Chutia, Tulika, and Nomi Baruah. "A review on emotion detection by using deep learning techniques." *Artificial Intelligence Review* 57.8 (2024): 203.

[28] Bender, Emily M., et al. "On the dangers of stochastic parrots: Can language models be too big?." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.

[29] Dominguez-Catena, Iris, Daniel Paternain, and Mikel Galar. "Metrics for dataset demographic bias: A case study on facial expression recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (2024): 5209-5226.

[30] Kamruzzaman, Mahammed, et al. "Seeing Race, Feeling Bias: Emotion Stereotyping in Multimodal Language Models." *Findings of the Empirical Methods in Natural Language Processing (EMNLP) 2025*. Association of Computer Linguistics, 2025.

[31] Mohammad, Saif. "Ethics sheet for automatic emotion recognition and sentiment analysis." *Computational Linguistics* 48.2 (2022): 239-278.

[32] Blease, Charlotte, and Adam Rodman. "Generative artificial intelligence in mental healthcare: an ethical evaluation." *Current Treatment Options in Psychiatry* 12.1 (2024): 5.

[33] Tavory, Tamar. "Regulating AI in mental health: ethics of care perspective." *JMIR Mental Health* 11.1 (2024): e58493.

[34] Dvorak, Fabian, et al. "Generative AI triggers welfare-reducing decisions in humans." *arXiv preprint arXiv:2401.12773* (2024).

[35] Ding, Yi, and Muzammil Najaf. "Interactivity, humanness, and trust: a psychological approach to AI chatbot adoption in e-commerce." *BMC psychology* 12.1 (2024): 595.

[36] Yang, Lei, and Shu Zhao. "AI-induced emotions in L2 education: Exploring EFL students' perceived emotions and regulation strategies." *Computers in Human Behavior* 159 (2024): 108337.

[37] Zhang, Heng, et al. "Emotional Artificial Intelligence in Education: A Systematic Review and Meta-Analysis." *Educational Psychology Review* 37.4 (2025): 106.

[38] Binz, Marcel, and Eric Schulz. "Turning large language models into cognitive models." *arXiv preprint arXiv:2306.03917* (2023).

[39] Pérez, Anxo, et al. "DepreSym: A Depression Symptom Annotated Corpus and the Role of Large Language Models as Assessors of Psychological Markers." *Language Resources and Evaluation* (2025): 1-26.

[40] Badawi, Abeer, et al. "Position: Beyond Assistance--Reimagining LLMs as Ethical and Adaptive Co-Creators in Mental Health Care." *arXiv preprint arXiv:2503.16456* (2025).

[41] Yu, Liangyue, et al. "Multimodal sensing-enabled large language models for automated emotional regulation: a review of current technologies, opportunities, and challenges." *Sensors* 25.15 (2025): 4763.

[42] Mishra, Riya, Braj Bhushan, and K. S. Venkatesh. "Toward Cross-Cultural Emotion Detection: The Indian Spontaneous Micro-Expression Dataset (ISMED)." *Frontiers in Psychology* 16 (2025): 1656104.

[43] Herrera-Poyatos, David, et al. "An overview of model uncertainty and variability in LLM-based sentiment analysis. Challenges, mitigation strategies and the role of explainability." *arXiv preprint arXiv:2504.04462* (2025).

[44] Lin, Yi-Cheng, et al. "Emo-bias: A large scale evaluation of social bias on speech emotion recognition." *arXiv preprint arXiv:2406.05065* (2024).

[45] Chhua, Kaylee, et al. "From Bias to Balance: Detecting Facial Expression Recognition Biases in Large Multimodal Foundation Models." *arXiv preprint arXiv:2408.14842* (2024).

[46] Lin, Zichao, et al. "Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models." *Artificial Intelligence Review* 57.9 (2024): 243.

[47] Pan, Bei, et al. "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods." *Neurocomputing* 561 (2023): 126866.

[48] Haque, MD Romael, and Sabirat Rubya. "An overview of chatbot-based mobile mental health apps: insights from app description and user reviews." *JMIR mHealth and uHealth* 11.1 (2023): e44838.

[49] Rzadeczka, Marcin, et al. "The efficacy of conversational artificial intelligence in rectifying the theory of mind and autonomy biases: Comparative analysis." *arXiv preprint arXiv* 2406 (2024).

[50] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

[51] Skjuve, Marita, Asbjørn Følstad, and Petter Bae Brandtzæg. "A longitudinal study of self-disclosure in human–chatbot relationships." *Interacting with Computers* 35.1 (2023): 24-39.

[52] Karkosz, Stanisław, et al. "Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial." *JMIR formative research* 8.1 (2024): e47960.

[53] Zhong, Wenjun, Jianghua Luo, and Hong Zhang. "The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis." *Journal of affective disorders* 356 (2024): 459-469.

[54] MacNeill, A. Luke, Shelley Doucet, and Alison Luke. "Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial." *JMIR Formative Research* 8 (2024): e50025.

[55] Kuhlmeier, Florian Onur, et al. "Combining Artificial Users and Psychotherapist Assessment to Evaluate Large Language Model-based Mental Health Chatbots." *arXiv preprint arXiv:2503.21540* (2025).

[56] Balcombe, Luke. "AI chatbots in digital mental health." *Informatics*. Vol. 10. No. 4. MDPI, 2023.

[57] Lee, Yoon Kyung, et al. "Large language models produce responses perceived to be empathic." *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2024.

[58] *He, Yuhao, et al. "Mental health chatbot for young adults with depressive symptoms: a single-blind, three-arm, randomized controlled trial." Journal of Medical Internet Research 24.11 (2022).*

[59] Sanjeewa, Ruvini, et al. "Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review." *JMIR Mental Health* 11 (2024): e58974.

[60] Hua, Yan Cathy, et al. "A systematic review of aspect-based sentiment analysis: domains, methods, and trends." *Artificial Intelligence Review* 57.11 (2024): 296.

[61] Bondarenko, Yuliia, et al. "Educational Milieu at Universities: Implementation of Inclusive Methods in High School Studies."

[62] Ajmani, Prerna, et al. "Patient behaviour analysis and social health predictions through IoMT." *2022 10th international conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO)*. IEEE, 2022.

[63] Wadhawan, Nisha, et al. "Leveraging Artificial Intelligence for Mental Health: A Comprehensive Review of Techniques and Applications." *International Conference on Data Science and Big Data Analysis*. Cham: Springer Nature Switzerland, 2025.