Contextual Emotion Detection in Mental Health Support Conversations Using Large Language Models

Florian Butler
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
fbutler@uab.edu

Rajesh Gaxena
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
rajesh2002@ucf.edu

Abstract

Mental health support increasingly takes place through text-based platforms, including crisis services, telepsychology portals, peer support forums, and mobile counselling applications. In these settings, emotional states must be inferred from written language, often across multiple conversational turns and under conditions of significant ambiguity. Contextual emotion detection therefore requires systems that can interpret not only discrete words but also dialogue history, user goals, social context, and clinical risk. Large language models offer substantial potential for this task because they encode rich linguistic context and can adapt to complex mental health expressions. However, their deployment in mental health infrastructures raises structural, ethical, and operational challenges that extend well beyond model accuracy. This paper presents a system-level analysis of contextual emotion detection in mental health support conversations using large language models. It examines architectural configurations, representation choices, model adaptation, deployment trade-offs, robustness, fairness, governance, sustainability, and clinical integration. The discussion emphasizes that effective emotion detection in mental health is not a standalone classification problem but a socio-technical system design problem. It requires careful balancing of inference quality, latency, privacy, interpretability, and accountability. The paper argues for modular architectures with human oversight, strong documentation practices, and policy alignment. It further considers cross-domain comparisons, energy costs, and the implications of using general-purpose foundation models in sensitive clinical contexts. The analysis is intended to inform researchers, system designers, and policy stakeholders who seek to build safer and more sustainable mental health text analysis infrastructures.

Keywords

contextual emotion detection; large language models; mental health; natural language processing; system architecture; fairness; governance; clinical deployment.

1. Introduction

Mental health support increasingly occurs through text-based channels, including crisis lines, telepsychology platforms, peer support forums, and mobile counselling applications. In these interactions, emotional states are not directly observable, and clinicians or automated systems must infer them from linguistic signals embedded in short messages, long transcripts, and

multi-turn conversations. The ability to detect fear, hopelessness, anger, ambivalence, guarded optimism, or emotional numbness is therefore essential for triage, safety monitoring, therapeutic empathy, and outcome evaluation. Contextual representations from masked bidirectional language models have enabled more nuanced text understanding than earlier lexicon-based or bag-of-words approaches [1]. Large-scale autoregressive models subsequently demonstrated that language models can perform complex inference with limited labeled data [2]. The underlying self-attention mechanism had already provided a general sequence-to-sequence architecture for modeling long-range dependencies [3], and unsupervised multitask learning expanded the range of tasks that could be addressed without task-specific architectures [4].

The recognition of emotion by machines has a longer history in affective computing [5], while emotion detection in conversational settings has been studied as a distinct challenge because of inter-turn dependencies, speaker roles, and implicit emotional cues [6]. In mental health applications, these technical capabilities intersect with broader social and ethical concerns about natural language processing systems [7]. Contextual emotion detection in this domain is not simply a matter of classifying text into anger, sadness, fear, or joy. It requires understanding how a person expresses distress over time, how a support worker responds, when risk escalates, and how language varies across cultural and clinical contexts. This paper examines the problem from a systems perspective, focusing on structural trade-offs, architecture, governance, infrastructure, deployment, sustainability, robustness, fairness, and policy implications. The central argument is that large language models can improve contextual emotion detection in mental health support conversations only when they are embedded within carefully designed socio-technical systems rather than deployed as isolated classifiers.

2. Background and Conceptual Foundations

Despite their expressive power, large language models also raise documented risks. Bender et al. [8] argued that scale alone does not produce understanding and may amplify biased associations learned from large uncurated corpora. Weidinger et al. [9] classified ethical harms from language models into multiple categories, including discrimination, misinformation, privacy leakage, and psychological harm. These concerns are particularly acute in mental health settings, where incorrect emotion inference can influence clinical decisions, safety assessments, or user experience. At the same time, efficiency-oriented methods such as knowledge distillation [10] and low-rank adaptation [11] have become important for real-world deployment because they reduce memory and computational demands while retaining much of the contextual competence of larger models. Foundation models more generally illustrate the shift toward reusable, general-purpose representations that can be adapted across tasks and domains [12].

From a global health perspective, the World Health Organization has highlighted the need for accessible and safe mental health services, particularly where professional resources are limited [13]. This creates both an opportunity and an obligation for automated support systems. In parallel, advanced sentiment classification methods have incorporated dynamic adaptive attention and supervised contrastive learning to improve discrimination among nuanced emotional categories [14]. Such advances can inform mental health emotion detection, although they must be adapted to clinical dialogue rather than general text. The broader foundation model literature emphasizes that contextual competence emerges from scale and pretraining, but downstream safety depends on data curation, domain adaptation,

evaluation, and human oversight. These foundations motivate a system-level rather than model-only analysis.

3. System Architecture for Contextual Emotion Detection

A systems view of contextual emotion detection requires a layered architecture that separates conversation ingestion, contextual encoding, emotion inference, explanation generation, safety filtering, and human review. In such an architecture, the language model operates within a controlled pipeline rather than as an opaque endpoint. The first layer normalizes transcripts, anonymizes identifiers, and preserves the temporal order of utterances. The second layer constructs a conversational representation that may include speaker roles, timestamps, crisis markers, and discourse history. The third layer applies a pretrained or domain-adapted model to infer emotional states, including ambiguous or mixed emotions. The fourth layer produces calibrated confidence estimates and natural language explanations that can be audited by clinicians or moderators. A final layer applies safety rules and routes high-risk cases to human responders. The energy cost of inference for large models has become a central concern [15], and this architectural separation allows smaller models or distilled variants to handle routine cases while reserving larger models for uncertain or high-risk interactions.

Modularity is especially important in mental health contexts because requirements differ across crisis hotlines, ongoing therapy platforms, asynchronous peer support, and research datasets. A crisis service may prioritize low latency, high recall for suicidal ideation, and strict privacy, while a therapy platform may prioritize longitudinal emotion tracking, clinician interpretability, and integration with electronic health records. A modular architecture enables each setting to select different encoding strategies, model sizes, and safety thresholds without retraining the entire system. The architecture also supports asynchronous auditing, because every inference can be logged with model version, input representation, confidence, explanation, and subsequent human decision. This auditability is essential for accountability and for learning from failures.

4. Modeling Emotional Context with Large Language Models

Emotion theory has long distinguished between discrete categories and continuous dimensions. Russell's circumplex model treats affect along pleasure and arousal axes [16]. The pleasure arousal dominance framework extends this with a dominance dimension to capture feelings of control or submission [17]. Large language models can support both perspectives. Discrete classification into anger, sadness, fear, or hopelessness can be useful for triage and structured reporting. Dimensional scoring can be more informative for clinical formulations because it captures mixed states, gradual shifts, and emotion intensity. A well-designed system may combine categorical screening with dimensional analysis, using the language model to generate intermediate natural language interpretations before mapping them to structured labels.

Contextual modeling is especially valuable when emotions are expressed indirectly, such as through sarcasm, understatement, metaphor, or avoidance. A person may not say they are afraid, but their repeated questions about safety, their sudden withdrawal from a topic, or their use of absolutes can indicate anxiety. Large language models can capture these signals across multiple turns, whereas fixed lexicons often fail. However, context window limitations mean that very long conversations must be summarized, chunked, or compressed. Each strategy introduces information loss and potential distortion. Systems must therefore be explicit about

how much dialogue history is preserved, which turns are considered most relevant, and how summaries are generated. These choices affect both clinical validity and fairness.

5. Structural Trade-Offs in Deployment

Deployment decisions involve unavoidable trade-offs between inference quality, latency, cost, privacy, and explainability. Model cards provide a structured format for documenting intended use, evaluation results, and limitations [18]. Without such documentation, mental health institutions may adopt emotion detection systems without understanding their failure modes. Post hoc explanation techniques can illuminate which words or phrases influenced a prediction [19], but these explanations are often approximate and may not capture the full contextual reasoning of a large model. In clinical settings, explanations must be treated as assistive artifacts rather than definitive causal accounts.

Cloud-based inference offers easy access to large models but may increase data exposure and dependency on external providers. On-premise deployment improves privacy control but requires specialized infrastructure and limits model size. Compact models trained with distillation, parameter-efficient adaptation, or task-specific fine-tuning can meet many clinical requirements at lower cost. A hybrid deployment model may use a compact on-premise model for routine emotion detection and an external large model only for difficult cases with explicit consent. Such configurations reduce latency, energy consumption, and privacy risk while preserving contextual depth where it is most needed. These trade-offs cannot be resolved by benchmark performance alone; they require careful analysis of clinical workflow, user expectations, and regulatory constraints.

6. Fairness, Robustness, and Governance

Fairness in mental health emotion detection requires attention to demographic variation, linguistic diversity, cultural expression, and mental health stigma. Ensuring fairness in machine learning to advance health equity demands that models be evaluated across patient populations and not only on aggregate metrics [20]. A model trained primarily on English-language, Western, or clinically referred populations may misread emotional expression in other groups. Datasheets for datasets provide a framework for documenting the provenance, composition, and potential biases of training data [21]. In mental health contexts, datasheets should also describe how diagnoses were assigned, whether crisis samples are included, and how consent was obtained. These practices improve transparency and support institutional review.

Robustness is another governance concern. Emotion detection systems may encounter adversarial or coercive language, ambiguous negation, slang, code words, or deliberate manipulation. A robust system should maintain stable behavior under such conditions and should flag uncertainty rather than silently producing confident predictions. Governance mechanisms should include periodic re-evaluation, incident reporting, clinician feedback loops, and clear procedures for overriding automated outputs. Human oversight must be substantive rather than symbolic, especially when emotion predictions inform risk assessment or treatment decisions. The design of governance structures is as important as the design of model architectures.

7. Infrastructure and Sustainability

The environmental impact of large language models has been widely documented, particularly in terms of training and inference energy costs [15]. Lightweight lexicon-based

systems such as VADER remain useful as baselines and preprocessing filters [22], but they cannot capture contextual nuance. Sustainable infrastructure therefore relies on a spectrum of models, from rule-based filters and distilled transformers to large models reserved for high-uncertainty cases. Distillation reduces model size and inference cost [10], while parameter-efficient adaptation lowers the cost of domain updating [11]. These methods help mental health organizations maintain local control without requiring continuous access to expensive external computation.

Sustainability also includes organizational sustainability. A system that requires frequent manual annotation, constant model retraining, and specialized engineering support may not be viable for community mental health providers. Therefore, deployment should emphasize reusable evaluation pipelines, clear documentation, and low-cost retraining cycles. Monitoring infrastructure should track drift in language use, changing slang, new crisis topics, and shifts in service populations. It should also track the proportion of cases escalated to humans, the frequency of model overrides, and the distribution of inferred emotions across groups. These operational metrics are essential for continuous improvement and for detecting emerging inequities.

8. Policy Implications and Clinical Integration

Policy frameworks for emotion detection in mental health must address privacy, consent, clinical safety, algorithmic accountability, and professional responsibility. General data protection regulations apply to mental health data with heightened sensitivity, but sector-specific guidance is often lacking. Clinical integration should be grounded in the principle that automated emotion detection is a decision support tool, not a replacement for clinical judgment. This principle should be embedded in system design, clinician training, and institutional protocols. The World Health Organization's emphasis on accessible mental health care [13] does not justify lowering safety standards through unregulated automation. Instead, it supports the use of well-governed systems to extend trained human capacity in resource-constrained environments.

Regulators and professional bodies should encourage model documentation, dataset transparency, and third-party auditing. Model cards [18] and datasheets [21] are practical instruments for this purpose. They allow mental health providers to compare systems, understand intended use, and assess whether a model is appropriate for their population. In addition, adverse event reporting should include both clinical and technical aspects, such as missed crises, false alarms, demographic disparities, and privacy incidents. Over time, such reporting can inform standards and improve safety. Policy should also address the procurement of commercial foundation models, requiring clear terms for data handling, model updates, and liability. These considerations are essential for public trust and for the long-term adoption of contextual emotion detection in mental health.

9. Conclusion

Contextual emotion detection in mental health support conversations is a promising but demanding application of large language models. The technical capability of these models to represent dialogue context, infer ambiguous emotions, and adapt to domain language is significant. However, their value depends on careful system design that addresses architectural modularity, fairness, robustness, energy consumption, privacy, and clinical integration. This paper has argued that mental health emotion detection should be approached as a socio-technical system rather than a standalone classification task. It requires layered

architectures with human oversight, structured documentation, continuous monitoring, and policy alignment. Modular deployment enables trade-offs between compact and large models, cloud and on-premise infrastructure, and categorical and dimensional emotion representations. Fairness and governance mechanisms must ensure that emotion detection does not amplify existing mental health inequities. Sustainability considerations demand attention to both environmental costs and organizational capacity. Future work should develop evaluation benchmarks that reflect clinical realities, study clinician interaction with model explanations, and examine longitudinal emotion dynamics across diverse populations. By addressing these structural dimensions, the field can build safer, more equitable, and more sustainable systems for supporting mental health through language technology.

References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics.

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

4. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report.

5. Picard, R. W. (1997). Affective computing. MIT Press.

6. Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Transactions on Affective Computing, 10(2), 243–257.

7. Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 591–598). Association for Computational Linguistics.

8. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery.

9. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Huang, P.-S., Ibarz, B., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

10. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

11. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

12. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

13. World Health Organization. (2022). World mental health report: Transforming mental health for all. World Health Organization.

14. Li, Q. (2026). Dynamic Adaptive Attention and Supervised Contrastive Learning: A Novel Hybrid Framework for Text Sentiment Classification. arXiv preprint arXiv:2604.10459.

15. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics.

16. Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

17. Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14(4), 261–292.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). Association for Computing Machinery.

19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). Association for Computing Machinery.

20. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

21. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

22. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225.