International Research Journal of Multidisciplinary Sciences ISSN:3107-930X(ONLINE)
VOL-1 ISSUE-9 Sept 2025 page:27-31

Emotion AI in the Era of Large Language Models: An NLP-Centric Survey

Anam Shariq Birla Public School, Doha Qatar anam.s.khan92@gmail.com

Abstract — The field of Affective Computing (AC), which aims to enable machines to recognize, interpret, and simulate human emotions, is undergoing a profound transformation driven by the advent of Large Language Models (LLMs). Traditionally reliant on multimodal signals (e.g., text, audio, visual) and specialized, task-specific models, AC is increasingly leveraging the rich world knowledge, contextual understanding, and generative capabilities of LLMs. This survey provides a comprehensive overview of the convergence of AC and LLMs from a Natural Language Processing (NLP) perspective. We first outline the traditional paradigms of affective computing in NLP. We then systematically review how LLMs are being utilized for emotion recognition, emotion reasonin g and explanation, and empathetic response generation. We delve into emerging techniques such as in-context learning, chain-of-thought prompting, and emotional alignment tuning. Furthermore, we critically examine the significant challenges and ethical considerations, including emotional bias in LLMs, the risk of anthropomorphism, and hallucination of emotional states. Finally, we propose future research directions, advocating for a more nuanced, culturally aware, and human-centric approach to building emotionally intelligent systems. This survey aims to serve as a foundational resource for researchers and practitioners at the intersection of NLP and affective computing.

Keywords — Affective Computing, Large Language Models, Emotion Recognition, Empathetic Response, Emotional Intelligence, Natural Language Processing, Survey.

1. Introduction

Human communication is inherently affective. Emotions color our language, shape our decisions, and form the bedrock of social interaction. For decades, the field of Affective Computing (AC) has sought to bridge the gap between human emotional experience and computational systems, with applications ranging from mental health support and educational technology to human-computer interaction and content recommendation [1]. A central pillar of AC is the processing of natural language, as text remains a primary medium for expressing and discerning emotional states.

Traditional NLP approaches to AC have largely relied on supervised learning with curated datasets. These methods include lexicon-based models using sentiment dictionaries (e.g., SentiWordNet) [2], and machine learning classifiers (e.g., SVMs, LSTMs) trained on labeled corpora to predict discrete emotion categories (e.g., joy, anger, sadness) or continuous dimensions (e.g., valence, arousal) [3]. While effective, these models often operate as "black boxes," lack genuine contextual understanding, and struggle with the complexity, ambiguity, and cultural specificity of emotional expression [4].

The rise of Large Language Models (LLMs) like GPT-4 [5], PaLM [6], and LLaMA [7] marks a paradigm shift. Pre-trained on vast and diverse corpora, LLMs have developed a remarkable capacity for semantic reasoning, commonsense understanding, and contextual nuance. This foundational capability presents an unprecedented opportunity to advance AC. An LLM can, in principle, understand that "This is just what I needed today!" could be sincere joy or bitter sarcasm, depending on the preceding conversation. It can generate not just emotionally appropriate but also contextually coherent and informative responses.

This survey explores this nascent and rapidly evolving landscape. We focus specifically on the NLP perspective, investigating how LLMs are being harnessed to understand and generate human emotion through language. We will:

Outline the transition from traditional AC methods to LLM-powered paradigms.

Survey key applications: emotion recognition, emotion explanation, and empathetic response generation.

Analyze the technical approaches, from zero-shot prompting to specialized fine-tuning.

Discuss critical challenges and ethical risks.

Propose promising future research directions.

2. From Traditional Models to LLM-Powered Affective Computing

The journey of AC in NLP has evolved through distinct stages. Early lexicon-based methods quantified emotion by matching words against pre-defined lists with associated sentiment or emotion scores [2]. While interpretable, they failed to capture context, negation, or sarcasm.

The era of supervised machine learning saw the adoption of models like Support Vector Machines (SVMs) and, later, Recurrent Neural Networks (RNNs) like LSTMs [3]. These models learned to map textual features (e.g., n-grams, word embeddings) to emotion labels from datasets like ISEAR [8] or GoEmotions [9]. Convolutional Neural Networks (CNNs) were also used to detect salient emotional phrases [10]. Despite their improved performance, these models were fundamentally limited by their training data, often failing to generalize to new domains or understand long-range contextual dependencies.

The introduction of pre-trained contextual embeddings like BERT [11] was a significant leap forward. Models fine-tuned on BERT could dynamically interpret word meaning based on context, leading to state-of-the-art results on many sentiment and emotion classification benchmarks [12]. However, these models were still primarily discriminative—excellent at labeling text but not at reasoning about emotion or generating empathetic dialogue.

The current LLM era builds upon this foundation but introduces qualitatively new capabilities. LLMs are not just feature extractors; they are generative, knowledge-rich, and highly adaptable reasoning engines. Their emergent abilities, such as incontext learning and chain-of-thought reasoning, allow them to perform affective tasks without task-specific training data, simply by following natural language instructions [13]. This shifts the paradigm from "training a model for emotion classification" to "instructing a general-purpose model to perform emotional understanding."

3. Key Applications and Techniques

3.1. Emotion Recognition and Classification

LLMs are being applied to emotion recognition in several innovative ways beyond simple fine-tuning.

Zero-Shot and Few-Shot Prompting: The simplest approach is to directly ask an LLM to identify the emotion in a given text. For example, a prompt like "Identify the primary emotion in the following text: '[User utterance]'. Choose from: joy, sadness, anger, fear, surprise, disgust." can yield impressive results, leveraging the model's embedded knowledge [14]. Few-shot prompting, providing a few examples, further improves performance and alignment with the desired label schema.

Explanation-Based Recognition: A more powerful technique leverages chain-of-thought (CoT) prompting [15]. Instead of a direct answer, the model is prompted to "reason step-by-step" (e.g., "The user says 'I just got the job!'. This is typically a positive event. The exclamation mark indicates excitement. Therefore, the emotion is likely joy."). This not only improves accuracy on complex examples but also provides a window into the model's "reasoning," enhancing interpretability [16].

Dimensional Emotion Analysis: Beyond categorical labels, LLMs can be prompted to analyze emotions along continuous dimensions like valence (pleasantness), arousal (intensity), and dominance [17]. For instance, an LLM can be asked to rate a statement on a scale of 1 to 9 for valence and arousal, demonstrating a more nuanced understanding.

3.2. Emotion Explanation and Cause Identification

Understanding that someone is sad is only the first step; understanding why they are sad is crucial for true empathy. LLMs excel at this task due to their extensive world knowledge. Given a conversational context, an LLM can be prompted to identify the likely causes of a user's emotional state [18]. For example, in a dialogue where a user expresses frustration after a computer crash, the LLM can infer the cause-and-effect relationship, something traditional models could not do. This capability is foundational for building advanced counseling and support systems.

3.3. Empathetic Response Generation

This is one of the most active and impactful application areas. The goal is to generate responses that acknowledge, understand, and react appropriately to a user's expressed emotion. Early chatbot responses were often generic ("I'm sorry to hear that"). LLMs enable a leap in quality.

Prompting for Empathy: Direct instructions like "Respond to the following message as a supportive and empathetic friend: [User message]" can generate highly tailored and context-aware responses [19].

Fine-Tuning on Counseling Corpora: LLMs can be fine-tuned on datasets of human-to-human empathetic conversations, such as counseling sessions [20] or platforms like Reddit's r/TrueOffMyChest, to learn the patterns and language of support [21]. This specializes the general-purpose LLM for the therapeutic domain.

Emotional Alignment Tuning: Drawing inspiration from Reinforcement Learning from Human Feedback (RLHF) used to align models like ChatGPT [5], researchers are exploring "Emotional RLHF." Here, human feedback is used to reward the model not just for being helpful and harmless, but for being genuinely empathetic and emotionally appropriate [22].

4. Critical Challenges and Ethical Considerations

The integration of LLMs into AC is not without significant challenges.

Emotional Bias and Stereotyping: LLMs trained on internet data can inherit and amplify societal biases. They may associate certain professions or demographics with specific emotions (e.g., consistently linking "nurse" with "compassion") or fail to understand the emotional expressions of marginalized groups [23]. This can lead to unfair and harmful interactions.

The Illusion of Empathy and Anthropomorphism: LLMs do not feel emotions; they simulate them based on statistical patterns. This risks creating a "illusion of empathy" where users form emotional attachments to a system that has no genuine understanding or care [24]. This raises ethical concerns, especially in vulnerable populations like children or those seeking mental health support.

Hallucination of Emotional State: LLMs can confidently generate incorrect analyses of a user's emotional state, a phenomenon known as "emotional hallucination" [25]. For example, they might infer sadness in a neutral statement. The consequences of such misdiagnosis in a therapeutic context could be severe.

Contextual and Cultural Nuance: While LLMs have broad knowledge, their understanding of culturally specific emotional expressions, sarcasm, and humor is still imperfect [26]. A joke in one culture might be an insult in another, and LLMs can easily misinterpret such subtleties.

Privacy and Manipulation: Affective systems have access to highly sensitive user data. The potential for emotional profiling and manipulation, for example in advertising or political campaigns, is a serious societal risk [27].

5. Future Research Directions

To address these challenges and realize the full potential of LLM-powered AC, we propose the following research directions:

1. Multimodal Affective LLMs: While this survey focuses on NLP, the future lies in integrating text with tone of voice (paralanguage) and facial expressions [28]. Developing LLMs that serve as a central reasoning engine for fused multimodal emotional signals is a key frontier.
2. Bias Mitigation and Fairness: Developing rigorous benchmarks and techniques to audit and debias LLMs for emotional tasks across different demographics is crucial [29]. This includes creating diverse and inclusive affective datasets.
3. Theory-Grounded and Psychologically Valid Models: Moving beyond simplistic emotion categories, future models should be grounded in robust psychological theories of emotion (e.g., Appraisal Theory [30]) and be validated against real-world human emotional experiences, not just textual annotations.
4. Explainable and Controllable Affective AI: Research is needed to make the emotional reasoning of LLMs more transparent and to allow users to control the "personality" and emotional style of the AI they interact with [31].
5. Long-Term Emotional Interaction: Most current work focuses on single-turn emotion recognition. A critical direction is modeling the dynamics of emotion over long-term conversations, tracking how emotional states shift and influence each other [32].
6. Robust Evaluation Frameworks: Moving beyond automated metrics like BLEU or F1-score, we need human-centric evaluation frameworks that assess perceived empathy, emotional supportiveness, and long-term user well-being [33].
7. Ethical Guidelines and Governance: The AC research community must proactively develop and adhere to strong ethical guidelines for the development and deployment of emotionally aware AI, with special protections for vulnerable users [34, 35].

6. Conclusion

The era of Large Language Models has irrevocably altered the landscape of Affective Computing. By endowing machines with unprecedented linguistic and reasoning capabilities, LLMs have enabled significant progress in emotion recognition, explanation, and empathetic response generation. Techniques like in-context learning and chain-of-thought prompting allow these models to perform nuanced affective tasks without explicit training, while fine-tuning and alignment methods can specialize them for supportive roles.

However, this power comes with profound responsibility. The risks of bias, illusion, and manipulation are real and must be addressed with rigorous research and thoughtful ethical practice. The path forward requires a collaborative effort between NLP researchers, psychologists, ethicists, and social scientists. The ultimate goal is not to create machines that feel, but to build tools that can better understand and respond to human emotion, thereby enhancing our well-being and enriching our interaction with technology. This survey has outlined the current state, the pressing challenges, and the promising future of this transformative convergence.

References

[1] R. W. Picard, Affective Computing . MIT Press, 1997.
[2] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in LREC , 2010.
[3] S. Poria et al., "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," Information Fusion , vol.
37, pp. 98–125, 2017.
[4] C. Strapparava and R. Mihalcea, "Learning to Identify Emotions in Text," in Proceedings of the 2008 ACM symposium on Applied computing , 2008.
[5] OpenAI, "GPT-4 Technical Report," 2023. [arXiv:2303.08774]
[6] R. Anil et al., "PaLM 2 Technical Report," 2023. [arXiv:2305.10403]
[7] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," 2023. [arXiv:2302.13971]

[8] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," Journal of personality and social psychology , vol. 66, no. 2, p. 310, 1994.
[9] D. Demszky et al., "GoEmotions: A Dataset of Fine-Grained Emotions," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , 2020.
[10] A. Severyn and A. Moschitti, "UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification," in Proceedings of SemEval-2015 , 2015.
[11] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT , 2019.
[12] S. M. Mohammad and F. Bravo-Marquez, "Emotion Intensities in Tweets," in Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (SEM 2017) , 2017.
[13] J. Wei et al., "Emergent Abilities of Large Language Models," 2022. [arXiv:2206.07682]
[14] T. Sun et al., "A Study of the Task and Model Agnostic Performance of Large Language Models," 2023. [arXiv:2305.16958]
[15] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in Advances in Neural Information Processing Systems , 2022.
[16] Z. J. Wang et al., "Chain-of-Thought Prompting for Understanding and Simulating Emotions in Text," in Findings of EMNLP , 2023.
[17] J. A. Russell, "A circumplex model of affect," Journal of personality and social psychology , vol. 39, no. 6, p. 1161, 1980.
[18] P. Zhong, et al., "Emotion Cause Analysis in Conversation with Large Language Models," 2023. [arXiv:2310.15570]
[19] R. Lowe et al., "Towards an Empathetic Chatbot that Knows When to Say 'I Don't Know'," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop , 2019.
[20] S. Sharma et al., "A Large-Scale Corpus for Counseling Conversation Research," in Proceedings of the 12th Language Resources and Evaluation Conference , 2020.
[21] A. See et al., "Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , 2019.
[22] Y. Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," 2022. [arXiv:2204.05862]
[23] S. L. Blodgett et al., "Language (Technology) is Power: A Critical Survey of "Bias" in NLP," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics , 2020.
[24] S. Weizenbaum, Computer Power and Human Reason: From Judgment to Calculation . W. H. Freeman and Company, 1976.
[25] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys , vol. 55, no. 12, pp. 1–38, 2023.
[26] L. Jiang et al., "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," 2023. [arXiv:2306.13063]
[27] B. C. Stahl and D. Wright, "Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation," IEEE Security & Privacy , vol. 16, no. 3, pp. 26-33, 2018.
[28] Y. Wang et al., "A Survey on Multimodal Large Language Models," 2023. [arXiv:2306.13549]
[29] N. Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," ACM Computing Surveys , vol. 54, no. 6, pp. 1–35, 2021.
[30] K. R. Scherer, "Appraisal Theory," in Handbook of cognition and emotion , T. Dalgleish and M. J. Power, Eds. John Wiley & Sons, 1999, pp. 637–663.
[31] H. Liu et al., "A Survey of Controllable Text Generation using Neural Networks," 2022. [arXiv:2201.05337]
[32] S. Poria et al., "DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition," in Proceedings of the AAAI Conference on Artificial Intelligence , 2021.
[33] C. P. Lee et al., "ROUGE: A Package for Automatic Evaluation of Summaries," in Text Summarization Branches Out , 2004.
[34] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems," First Edition, 2019.

[35] A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," Nature Machine Intelligence , vol. 1, no. 9, pp. 389–399, 2019.