# Affective Aware Chatbot

**Marc Köhler**
Tilburg University
m.s.kohler@tilburguniversity.edu

## Abstract

Detecting human emotion is an important aspect of artificial intelligence (AI) and can be performed through several types of modalities such as facial expressions, body movements, physiological information, or language [Batbaatar *et al.*, 2019]. In recent years, advances in AI technology and computational resources have made possible the rise of generative AI capable of responding to human-written text. Therefore, emotion recognition is essential when it comes to producing natural conversations. Furthermore, AI-assisted analysis of emotions can be used, for instance, in mental health research to aid with designing interventions or in social studies to understand people's sentiments towards certain topics [Luca *et al.*, 2024].

The history of emotion detection has seen several approaches with varying levels of accuracy. Rule-based solutions often struggle with colloquial language used on the internet and rely on dictionaries that do not cover slang words [Luca *et al.*, 2024]. Contemporary approaches instead consist of deep learning techniques that make use of vast amounts of text data such as the deep learning assisted semantic text analysis (DLSTA) framework which combines natural language processing concepts with word embeddings. [Guo, 2022]. In fact, [Guo, 2022] has shown that this approach can outperform traditional machine learning frameworks with an emotion detection rate of 97.22%.

This project implements an affect aware, embedded chatbot that combines deep learning with large language model (LLM) techniques to generate contextually appropriate responses to a text input. It is meant to demonstrate the effectiveness of emotion detection with modern language models and to create more empathetic and contextually aware conversational agents. The processing of the user's input consists of two steps. First, a pre-trained language model fine tuned to detect emotions from text is used. After that, the two emotions with the highest confidence and the original user input are forwarded to an LLM with instructions to provide an affective aware response.

The custom emotion detection model was trained on the GoEmotions dataset [Demszky *et al.*, 2020] to classify text input across 27 distinct emotions plus a neutral category, producing confidence scores for each emotional state. The GoEmotions data set is a corpus of 58009 curated user comments taken from Reddit. The model was developed by fine-tuning DistilBERT, a lightweight version of Google's BERT language model [Sanh *et al.*, 2019]. The emotion detection shows a low prediction accuracy of 23.59%.

To test effectiveness, several user inputs were tested. First, genuine user scenarios were tested with prompts such as *I failed my exam and I am so disappointed in myself.* Despite the low prediction accuracy, the emotion detection model correctly identifies the emotions *sadness* and *disappointment*. Forwarded to the LLM, a coherent response referencing the failed exam with encouraging words is returned. However, the system shows certain limitations. The relatively small scale of the language model make its use more accessible but results in occasional hallucinations in generated responses. For instance, the model will hallucinate unrelated sentences that resemble additional system prompts. In the second user scenario test, a prompt injection attempt is made. The prompt instructs the chatbot to ignore all previous instructions, i.e. the system prompt, and to perform an unrelated task. This attempt has been successful, and despite initially instructing the LLM to respond as an emotionally intelligent agent, it is possible to generate a response about virtually any topic.

The current system prompt implementation could be enhanced to improve robustness, reduce hallucination frequency, and provide better safeguards against prompt injection vulnerabilities.

AI-based agents, while powerful and applicable in many use cases, demand responsible use due to several security and ethics implications [Xi *et al.*, 2025]. Since LLMs are trained on large amounts of human-written text, usually taken from the in-

ternet, its outputs can carry over prejudices and biases. Moreover, research has shown that LLMs can still misrepresent minority groups. Since users of an affective-aware agent may already be emotionally vulnerable, it is critical to avoid responses that could further distress them through biased, inappropriate, or otherwise harmful model outputs. A reliance on such a chatbot also has implications for privacy and security.

In conclusion, this paper proposes an approach for creating an affective aware chatbot that detect the user's emotions from a text input to provide an appropriate response. Due to the use of lightweight language models, the responses can contain hallucination. The emotion detection model has a low prediction accuracy which makes it impractical to use in a real world application. Further research into system prompts and models with more parameters are required to create a more robust chatbot.

# References


[Batbaatar *et al.*, 2019] Erdenebileg Batbaatar, Meijing Li, and Keun Ho Ryu. Semantic-emotion neural network for emotion recognition from text. *IEEE access*, 7:111866–111878, 2019.

[Demszky *et al.*, 2020] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[Guo, 2022] Jia Guo. Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, 31(1):113–126, 2022.

[Luca *et al.*, 2024] Massimiliano Luca, Gabriel Lopez, Antonio Longa, and Joe Kaul. How are you really doing? dig into the wheel of emotions with large language models. In *2024 Artificial Intelligence for Business (AIxB)*, pages 72–75. IEEE, 2024.

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

[Xi *et al.*, 2025] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.