Emotion-Adaptive Multimodal CBT Assistant

An Applied Machine Learning System for Multimodal Emotion-Aware Interaction

Adhish Nanda, Duc Toan Huynh
MSc Data Science
University of Europe for Applied Sciences
Berlin, Germany

***Abstract*—This project presents an end-to-end multimodal machine learning system integrating emotion recognition from text, audio, and visual inputs with structured Cognitive Behavioural Therapy (CBT)-aligned response generation. A late-fusion architecture combines pretrained encoders with a lightweight task-trained fusion head. The system emphasizes modularity, reproducibility, and safe deployment principles while demonstrating applied ML techniques in multimodal affective computing.**

I. Project Overview

The system was designed to move beyond text-only emotion detection by integrating multimodal signals in a modular and interpretable pipeline. Key objectives included:

- Multimodal support (text, audio, visual)
- Clear separation of pretrained vs task-trained components
- Robust handling of missing modalities
- Emotion-aware CBT-style response generation

II. System Architecture

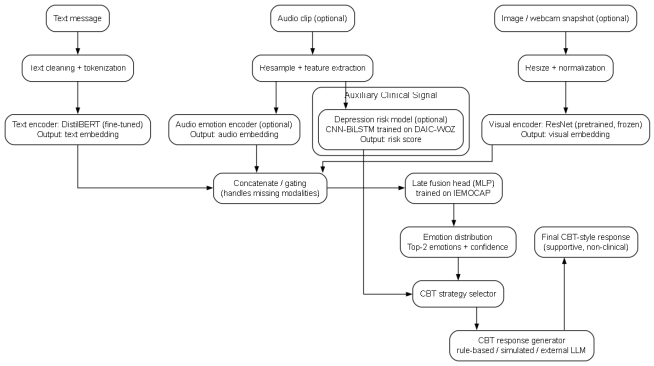

Fig. 1. Modular multimodal architecture with late fusion and CBT-aligned response generation.

Core components include: DistilBERT (text), ResNet (visual, frozen), CNN–BiLSTM (audio depression signal), and a lightweight MLP fusion head trained on IEMOCAP.

III. Datasets

- MELD – Conversational text emotion recognition
- IEMOCAP – Multimodal emotion recognition
- DAIC-WOZ – Speech-based depression interviews

TABLE I
Performance Summary

Model	Accuracy	Macro F1
Text (MELD)	54.1%	40.8
Text (IEMOCAP)	10.0%	6.6
Visual (ResNet, IEMOCAP)	13.6%	6.2
Audio Depression (DAIC)	65.7%	39.7

IV. Quantitative Results

Key Observations:

- MELD text model performs strongest (neutral F1 = 70.3%)
- IEMOCAP text-only setup underperforms (label/instability issues)
- Visual branch weak as standalone (frozen feature extractor)
- Depression model shows majority-class bias (0% F1 for depressed)
- System-level demo confirms effective multimodal emotional adaptation

V. System-Level Demonstration

Qualitative evaluation shows multimodal disagreement handling (e.g., neutral text/face but sadness in voice). The system adapts CBT-style responses accordingly, maintaining non-clinical and safety-aware framing.

VI. Engineering Decisions

- Late fusion chosen for interpretability and controlled experimentation
- Majority of encoders frozen to reduce overfitting and compute cost
- Modular structure enabling reproducibility and debugging
- Ethical framing integrated into response logic

VII. Technical Stack

Python, PyTorch, Transformers (DistilBERT), ResNet, CNN–BiLSTM, scikit-learn, OpenCV, NumPy, Pandas.

VIII. Ethical Framing

This system is a research prototype and not a diagnostic tool. Depression-risk output is used only as contextual modulation and does not replace professional care.

Source code and documentation available on GitHub.