Empathy-Driven AI: An Emotion-Responsive Companion for Elderly Care

Tanvi Bhaskarwar
bhaskarw@usc.edu

Saptarshi Mondal
saptarsh@usc.edu

Mayank Patil
mmpatil@usc.edu

Abstract

This research extends previous work on multimodal sentiment analysis by integrating enhanced models and a fusion methodology to unify visual and audio modalities. Using the CMU-MOSEI dataset, we demonstrate a comprehensive sentiment prediction framework achieving accuracy rates of 80-85%. Additionally, a response generation module inspired by the Emotional Chain-of-Thought (ECoT) model enables empathetic, contextually appropriate responses. This work advances the field of affective computing by establishing a robust framework for sentiment analysis based on facial expressions and audio features, with potential applications in human-computer interaction systems.

1 Introduction

Social isolation among elderly individuals poses significant mental health risks, including depression and cognitive decline. This project aims to develop an empathetic AI companion capable of detecting emotional states through facial and vocal cues while generating supportive responses. Key contributions include feature alignment, fusion of multi-modal data, and a response generation system inspired by ECoT.

Challenges Addressed

1. **Data Complexity:** Temporal synchronization of CMU-MOSEI's multimodal datasets.
2. **Model Scalability:** Optimized neural architectures balancing computational efficiency and prediction accuracy.
3. **Fusion Implementation:** Hybrid attention-based model combining visual and audio features.
4. **Empathetic Responses:** Real-time response generation using ECoT-inspired prompts.

2 Related Work

- **Our method builds upon three key existing** approaches:
  1. Artificial Emotional Intelligence in Socially Assistive Robots [1]
     - **Their Approach:** [1] Uses facial expression recognition (FER) and sentiment analysis with weighted averaging
     - **Our Implementation:**

Preprint. Under review.

        * Similarly uses multi-modal analysis but focuses on CMU-MOSEI dataset
        * Extended their approach[4] by combining VisualFacet42,OpenFace and COVAREP features
        * Implemented more efficient batch processing (16 samples fo video or 256 samples for audio) along with using models that are more efficent with large datasets (such as XGBoost) for real-time analysis
    2. Facial Emotion Recognition Using Video and Audio[3]
        - **Their Approach:[3]** Uses ResNet50V2, ResNet152V2 for visual processing and CNNs as well as MLPs for audio dataset
        - **Our Implementation:**
            * Adopted similar ResNet architecture but optimized for our high-dimensional feature space (11,968 dimensions)
            * Enhanced with custom MLP for better processing of combined features
            * Experimented with Neural Networks, XGBoost, CNN, SVR and Tree Regressor for audio dataset. The models were optimized for the high dimensional dataset (73 features)

- **Strengths and Weaknesses Analysis**
    1. **Strengths of Existing Methods:**
        - Real-time processing capability (87.7% accuracy)
        - Multi-modal integration techniques
        - Established model architectures (ResNet, CNN)
    2. **Key Weaknesses and Our Solutions:**
        - **Resource Limitations:**
            * Original Weakness: Heavy computational requirements
            * Our Solution:
                · Implemented efficient batch processing with respect to model
                · Optimized memory usage through careful feature selection
                · Developed streamlined pre-processing pipeline
        - **Data Integration:**
            * Original Weakness: Limited feature alignment between modalities
            * Our Solution:
                · Created custom alignment methodology for VisualFacet42, COVAREP and OpenFace datasets
                · Implemented synchronization techniques for temporal consistency
    3. **Elderly-Specific Challenges:**
        - Original Weakness: Models not optimized for elderly expressions
        - Our Solution:
            * Focus on robust feature extraction suitable for varied age groups
            * Enhanced preprocessing to handle different speech patterns

- **Evaluation Methods and Results**
    1. Benchmark Comparison:

Category	Metric	Dataset/Context	Result
Previous Work	Accuracy	AffectNet dataset	80%
	Real-time accuracy	-	87.7%
	F1 Score	Emotion prediction	76.59%
Our Results	Accuracy	Test data	80-85%
	Evaluation Metrics	Cross-validation on CMU-MOSEI dataset	Performance stability across different expressions and resource efficiency metrics

Table 1: Comparison of Previous Work and Our Results

2. The results are convincing because:
   - Comparable accuracy to existing methods
   - More efficient resource utilization
   - Better handling of multi-modal feature alignment
   - Specific focus on elderly interaction patterns

3 Methods

- **Methodology**
  This project tackles sentiment analysis based on facial expressions and audio cues by creating a pipeline for feature extraction and classification. Visual features from the CMU_MOSEI VisualFacet42, COVAREP and OpenFace 2.0 datasets are aligned, merged, and input into custom models for predicting sentiment.

  1. Feature Extraction and Dataset Processing:

     CMU-MOSEI Dataset Processing:
     * Synchronization of **VisualFacet42** (42 facial features) and **OpenFace 2.0** facial landmarks
     * Custom **VisualSentimentDataset** class implementation for efficient data handling
     * Feature extraction of **COVAREP** dataset (17 million frames, 73 features)
     * This dataset is combined with the sentiment-emotion output to produce a custom AudioSentiment dataset for efficient data handling
     * **Feature alignment** to ensure temporal consistency across modalities and consistency with the sentiment and emotion outputs
     * **Data normalization** and standardization for consistent model inputs

  2. Model Architecture and Feature Processing:

     Visual Feature Processing:
     * VGG16 implementation for initial facial feature extraction
       · Pretrained on ImageNet for robust feature recognition
       · Fine-tuned on facial expression data
       · Extraction of high-level visual representations from video frames

     Custom Neural Network Design:
     * Multi-Layer Perceptron (MLP) Architecture:
       · **Input layer:** Dimensioned to accept combined features (11,968 dimensions)
       · Hidden layers: Multiple fully connected layers with varying sizes
       · **ReLU activation** functions for non-linear transformations
       · **Dropout layers** (rate: 0.5) for regularization
       · **Output layer:** Softmax activation for sentiment classification

     Custom Audio models
     * Multi-task Neural Network:
       · **Input layer**: Dimensioned to accept combined features(73 dimensions)
       · **Hidden Layers**: 2 fully connected Layers with 128 and 64 neurons respectively
       · **ReLU activation** functions for non-linear transformation
       · **Dropout Layers** (rate 0.4 and 0.3) to prevent overfitting
       · **Output Layer:** Tanh activation function for sentiment output and relu activation fuction for emotion output
     * XGBoost:
       · **Model Architecture**: XGBRegressor with **reg:squarederror** as the objective and the tree-method is **gpu-hist**
       · **Parameters**: The general parameters used for the **sub-sample** was either 0.6, 0.8 or 1 for any of the emotions, the **n-estimators** was either 100, 200 or 500, the **max-depth** was 10 or 8, the **learning-rate** was 0.1 or 0.2 and the **colsample-bytree** was either 1.0, 0.5 or 0.7. Each emotion took one of the values for each of the mentioned parameters. They could be the same values or different.

* We also built Tree Regressor, SVR,CatBoost,LightGBM,AdaBoost and CNN models which were very primitive and basic in nature and did not give the best results and were not included.CatBoost and LightGBM use all default values with verbose 1.

3. Training Infrastructure

   Data Management:
   * **Dataset splitting:** 80% training, 20% testing
   * **Mini-Batch Gradient Descent** for faster convergence and model efficiency.
   * **Adam Optimizer** with a learning rate of 0.001 for video and 0.0005 for audio to achieve adaptive learning and effective convergence.

   PyTorch DataLoader implementation for Video dataset:
   * **Batch size:** 16 samples for video
   * Shuffle enabled for training set
   * **Mini-batch** gradient descent optimization

   Tensorflow keras used for Audio dataset Neural network:
   * **Batch size:** 256 samples for audio
   * Shuffle enabled for training set
   * **Mini-batch** gradient descent optimization

   XGBoost Configuration:
   * No Batch size required
   * Hyperparamter tuning for optimal model performance with respect to each emotion
   * Cross-validation for assessing model stability and performance

   Training Configuration:
   * **Loss function:** Cross-entropy loss for multi-class classification on Video dataset Mean Squared Error for audio dataset
   * **Optimizer:** Adam with learning rate 0.001 and 0.0005
   * Batch normalization for stable training
   * Early stopping to prevent overfitting
   * Learning rate scheduling for optimization

4. Implementation Details

   P**yTorch Framework:**
   * Custom dataset class extending torch.utils.data.Dataset
   * Efficient GPU utilization for accelerated training
   * Vectorized operations for feature processing

   Performance Optimization:
   * Memory-efficient batch processing
   * Parallel data loading with multiple workers
   * Gradient accumulation for large feature spaces

5. Fusion Model
   - **Architecture:** The Attention-based Fusion Model combines embeddings from visual and audio modalities into a shared latent space. Attention weights are dynamically calculated to highlight the most relevant features, improving classification performance.
   - **Why Attention-based Fusion?:** Compared to simple concatenation or summation fusion methods, attention allows the model to assign different importance levels to modalities based on the context. This is especially useful when one modality (e.g., audio) provides more reliable cues for specific emotions.
   - **Comparison to Other Techniques:**
     * **Concatenation:** Involves combining modality embeddings directly. While simple, it treats all modalities equally, which can lead to suboptimal performance in imbalanced scenarios.
     * **Weighted Average**: Assigns fixed weights to each modality. This approach lacks adaptability to varying contexts.
     * **Attention-based Fusion:** Dynamically adjusts modality contributions based on context, improving both accuracy and interpretability.

  - Training Configuration:
    * **Optimizer:** Adam with learning rates of 0.001 (visual) and 0.0005 (audio).
    * **Loss Functions:**
      · Cross-entropy for multi-class sentiment classification.
      · Mean squared error for continuous emotion metrics.
    * **Techniques:** Early stopping, batch normalization, and dropout layers to prevent overfitting.
  - **Fusion Layer:** Computes attention weights to aggregate features dynamically, ensuring context-dependent importance.
  - **Classifier:** Final dense layers predict sentiment and emotion.
  - Performance improvements from fusion were quantified by comparing single-modality and fused-modality results, showing up to 10% higher accuracy.
- The model's performance is evaluated on a distinct test set, measuring metrics such as accuracy, F1 score, and a detailed classification report. In our use case, a prediction is termed as true positive, if it can correctly predict the presence of the particular emotion for that record, and false negative if a particular emotion is not predicted. All the outputs use 5-fold cross validation. By using separate training and test sets, we ensure that the model's effectiveness remains generalizable and not over-fitted to the training data. Additionally, we conduct validation through experiments with different model architectures and hyper-parameters, including variations in layer count and dropout rates, to identify the optimal configuration for peak performance.
- **Potential Challenges and Risks**
  - Overfitting Risk:
    * Limited feature sets for certain expressions can lead to overfitting, especially with complex model architectures.
    * To address this, dropout layers and regularization techniques are employed to enhance generalization.
  - Data Alignment and Processing Challenges:
    * Aligning features across datasets is complex, as feature arrays vary in length across videos.
    * Adjustments to feature preprocessing may be necessary to ensure compatibility and consistency among datasets.
  - Hardware Limitations:
    * Training on large datasets requires substantial computational power, making memory and training time potential bottlenecks.
    * Efficient batching and possible model simplifications are used to manage these limitations.

4 Experiments

Training Progress: Throughout the early training epochs, the loss consistently decreased, demonstrating that the models were learning effectively. The average loss for each epoch was recorded for every batch, highlighting gradual advancement.

Model Accuracy: During the evaluation phase, accuracy of the model was assessed on the set-aside test data. The trained model displayed encouraging accuracy rates, with findings suggesting accuracy levels of about 80-85% on the test data, varying by specific model and settings.

Data Inspection: Thorough inspections were performed on feature shapes and distributions to confirm compatibility with the models. Techniques for reshaping and flattening were experimented with to ensure data conformed to model input standards.

For the audio dataset, we built 9 models: Tree Regressor, a SVR, a multilayer neural network, xgboost,AdaBoost,XGBoost,LightGBM,CatBoost and finally a convolution neural network. The

metrics that we used to determine how well the models worked are given in the table along with the results and only the models with the best outpus have been displayed. All the results were done using cross validation

Model Name	Accuracy	F1 Score	Precision	Recall
Multilayer Neural Network	0.78 ± 0.02	0.10 ± 0.01	0.46 ± 0.03	0.09 ± 0.01
XGBoost	0.80 ± 0.01	0.16 ± 0.02	0.86 ± 0.01	0.13 ± 0.01
AdaBoost	0.68 ± 0.17	0.66 ± 0.23	0.76 ± 0.18	0.68 ± 0.17
LightGBM	0.83 ± 0.14	0.79 ± 0.14	0.82 ± 0.13	0.83 ± 0.14
CatBoost	0.83 ± 0.14	0.80 ± 0.13	0.82 ± 0.13	0.83 ± 0.14

Table 2: Model performance metrics with standard deviations on audio dataset

Model Name	Accuracy	F1 Score	Precision	Recall
ResNet-34	0.7252 ± 0.13	0.6097 ± 0.13	0.5260 ± 0.12	0.7252 ± 0.13
VGG-16	0.830 ± 0.16	0.800 ± 0.22	0.7700 ± 0.17	0.8300 ± 0.16
TimeSformer (SOTA)	0.8100 ± 0.13	0.7800 ± 0.13	0.7500 ± 0.12	0.8200 ± 0.13

Table 3: Model performance metrics of overall emotion on video dataset

Model Name	Accuracy	F1 Score	Precision	Recall
Attention-based Fusion	0.850 ± 0.17	0.8312 ± 0.13	0.830 ± 0.16	0.830 ± 0.15

Table 4: Fusion Model

5 Response Generation

Overview
The response generation component uses the Emotional Chain-of-Thought (ECoT)[2] method to craft empathetic and context-aware replies. A fine-tuned Large Language Model (LLM) and LangChain framework ensure alignment with human emotional intelligence principles.

Fine-Tuning and LangChain Integration
The LLM was fine-tuned on emotionally annotated datasets (e.g., IEMOCAP, DailyDialog) to improve emotion recognition and response generation. LangChain structured the workflow:

- **Prompt Templates:** ECoT-based prompts guided the LLM to generate emotionally intelligent responses.
- **LLM Chains:** Modular pipelines connected emotion detection, prompts, and generation.

Emotional Chain-of-Thought (ECoT)
ECoT, inspired by Goleman's Emotional Intelligence Theory, follows four steps:

1. Understanding Context: Analyze user input for emotional cues.
2. Recognizing Emotions: Identify user emotions (e.g., sadness, joy, surprise).
3. Regulating Emotions: Ensure responses are empathetic and aligned.
4. Influencing Emotions: Provide supportive and constructive replies.

Example:

- **Input Emotion: Surprise**
  Generated Response: "It sounds like something unexpected has caught you off guard! Take a moment to process it—surprises can bring opportunities too. How can I support you right now?"
- **Input Emotion: Sadness**
  Generated Response: "I'm sorry you're feeling down. It's okay to take a moment for yourself. Remember, small acts of self-care can make a difference—can I help you brainstorm some ideas?"

Results
The system achieved higher Emotional Generation Scores (EGS) compared to baseline methods, demonstrating significant improvements in emotional alignment and response coherence. ECoT and LangChain helped break down tasks into manageable steps, enhancing the empathy and utility of the responses.

6 Discussion

- The experiments conducted so far indicate that both model architectures can successfully process visual sentiment features along with audio sentiment and emotion features to make accurate predictions on facial expression-based sentiment labels and audio-based sentiment/emotion labels. The consistent decrease in training loss and promising accuracy on the test set (approximately 80-85%) suggest that the models are generalizing well to unseen data. This aligns with our goal of achieving a baseline accuracy that reflects a meaningful understanding of sentiment in facial expressions.
- A key insight from the results is the potential of the fusion model to integrate audio and visual data predictions. By combining emotions detected from both modalities, the fusion model can generate a more accurate final emotion prediction. This multimodal approach leverages the strengths of both audio and visual data, leading to improved sentiment and emotion recognition even in complex or ambiguous cases where a single modality may fall short.
- Beyond emotion classification, the fusion model's output can be utilized for response generation. Specifically, the detected emotion can inform the generation of responses aimed at alleviating distress or regulating negative emotions. For example, when the model detects signs of frustration or sadness, the response generation system can provide empathetic feedback or calming suggestions, enhancing human-computer interaction. This capability is particularly useful in applications such as mental health support systems, virtual assistants, or educational tools where emotional regulation plays a critical role.
- The outcomes achieved up to this point align closely with our initial projections. We expected visual sentiment and audio sentiment classification to achieve moderate to high levels of accuracy, considering the utilization of established models and preprocessing methods. However, we also anticipated the potential for overfitting, particularly due to the high-dimensional feature inputs. Preliminary findings indicate that dropout and other regularization techniques are effectively reducing overfitting, which corresponds with our expectations regarding model resilience.
- To confirm the findings so far, conducting additional experiments with different configurations was crucial. Increasing the number of training epochs and fine-tuning hyperparameters like learning rates and dropout rates shed light on the models' consistency and performance during prolonged training. Furthermore, assessing the models across various subsets of the dataset helped ascertain whether the results were applicable across a range of facial expressions and conditions.

References

[1] H. Abdollahi, M. H. Mahoor, R. Zandie, J. Siewierski, and S. H. Qualls. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, 14(3):2020–2032, Jul 2023. doi: 10.1109/taffc.2022.3143803.

[2] Z. Li, G. Chen, R. Shao, Y. Xie, D. Jiang, and L. Nie. Enhancing emotional generation capability of large language models via emotional chain-of-thought, Aug 2024. URL `https://arxiv.org/abs/2401.06836`.

[3] A. Shrivastava, D. Dubey, M. Verma, and H. Verma. Facial emotion recognition using video and audio. *International Journal of Research Publication and Reviews*, 5(1):2517–2527, Jan 2024. doi: 10.55248/gengpi.5.0124.0261.

[4] R. Zandie and M. H. Mahoor. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *Computing Research Repository*, abs/2003.02958, 2020.