Project Document: Multimodal Emotion Recognition System

(Text, Audio, and Video/Faces)

1. Project Overview

The **Multimodal Emotion Recognition System** is an AI-powered model designed to detect human emotions from **text**, **voice**, and **video-based facial expressions**. The system uses advanced pre-trained deep learning models that have been fine-tuned using publicly available and open-source datasets.

By combining multiple input channels, the system can identify emotional states such as:

- **Happy**
- **Sad**
- **Angry**
- **Neutral**
- **Fear**
- **Surprise**

This makes the system highly suitable for applications like:

- Customer support
- Interview analysis
- Mental health monitoring
- Student engagement monitoring
- Human–computer interaction

---

2. System Architecture

The system consists of **three independent emotion detection modules**, each handling one modality. Their outputs are then merged for overall emotion prediction.

User Input → Preprocessing → Feature Extraction → Prediction → Emotion Output

a) Text Emotion Recognition Module

- Identifies emotions from sentences or paragraphs.
- Uses a multilingual AI model trained on publicly available global datasets.
- Supports **English and Tamil** language inputs.
- Capable of classifying at least 5–7 emotional categories.

b) Audio Emotion Recognition Module

- Takes raw audio or speech and extracts emotional cues from:
    - Pitch
    - Tone
    - Energy
    - Mel-frequency cepstral coefficients (MFCCs)
- Uses pre-trained speech emotion datasets for training and fine-tuning.
- Works with English and Tamil speech.

c) Video/Facial Emotion Recognition Module

- Detects facial expressions from video frames.
- Uses lightweight CNN or transformer-based models.
- Recognizes facial cues such as:

    - Smile
    - Frown
    - Eyebrow movement
    - Eye openness
- Supports real-time emotion detection from webcam or video files.

---

3. Datasets Used (Open-Source)

This project does **not use any copyrighted or private data**.
It is trained and evaluated using **open datasets**, such as:

Text Datasets

- **GoEmotions Dataset** (Google)
- **TamilEmo Dataset** (for Tamil emotion sentences)

Audio Datasets

- **IEMOCAP Speech Emotion Dataset**
- **RAVDESS Emotional Speech Dataset**

Video/Faces Datasets

- **FER-2013 Facial Emotion Dataset**
- **CK+ Facial Expression Dataset**
- **AffectNet** (subset available publicly)

These datasets provide diverse samples that help the model understand both **linguistic** and **visual** emotional patterns.

---

4. Features & Capabilities

✔ Multilingual Emotion Recognition

- Tamil & English for text and voice

✔ Real-Time Video Emotion Tracking

- Detect emotions frame-by-frame
- Suitable for interviews, online classes, meetings

✔ Audio-Based Emotion Analysis

- Works with microphone input or uploaded audio files

✔ Text Emotion Analysis

- Instant emotion detection from any written message

✔ Fusion Logic

- When all three inputs are available, a **weighted fusion technique** is used to determine the overall emotion.

---

5. Technology Stack

- **Python**
- **Transformers (HuggingFace)**
- **PyTorch / TensorFlow**

- **OpenCV (for video & face detection)**
- **Librosa (audio processing)**
- **Pre-trained multilingual Transformer models**
- **Lightweight CNN for FER (Facial Emotion Recognition)**

---

6. Workflow Diagram

Step 1: Input

- User enters text
- User records or uploads audio
- User provides video/webcam feed

Step 2: Preprocessing

- Text tokenization
- Audio feature extraction
- Face detection in frames

Step 3: Emotion Classification

- Text model → emotion label
- Audio model → emotion label
- Video model → emotion label

Step 4: Fusion

The system merges all outputs using:

- Majority voting
- Confidence averaging
- Attention-based weighting (optional)

Step 5: Final Output

The system returns:

- Primary emotion
- Confidence score
- Per-channel prediction
- Optional visualization (graphs)

---

7. Use Case Scenarios

🎓 Education

Student engagement monitoring in online classrooms.

🧑‍💼 HR & Interviews

Emotion tracking during interviews to understand candidate confidence and stress.

☎ Customer Support

Understanding customer sentiment in calls or chat.

🏥 Healthcare

Monitoring emotional well-being for mental health support systems.

---

8. Advantages of This System

1. **Multimodal understanding**
   More accurate than single-modality emotion systems.

2. **Language flexibility**
   Supports Tamil + English.

3. **Real-time capability**
   Works with webcam and microphone.

4. **Fully offline deployment**
   Can run locally without any cloud dependency.

5. **Pre-trained & fine-tuned models**
   Fast processing and high accuracy.

---

9. Limitations

- Accuracy depends on clarity of audio and lighting condition in video.
- Cultural differences in expressions may affect results slightly.
- Multimodal fusion increases computational requirements.

---

10. Future Improvements

- Add more languages (Hindi, Telugu, Spanish, Arabic).
- Integrate body-gesture emotion analysis.
- Implement transformer-based multimodal fusion (e.g., FLAVA, MMBT).
- Deploy as a web dashboard or mobile app.

---

11. Conclusion

The Multimodal Emotion Recognition System provides a **complete AI solution** for detecting emotions from text, speech, and facial expressions using pre-trained, publicly available datasets.

It performs efficiently in real-time and supports Tamil and English, making it highly applicable in Indian academic, corporate, and customer service environments.

This solution demonstrates modern AI capability and can be customized further based on client needs.