MuseMood:

A Dynamic Multimodal System for Robust Input Fusion

Demo by **Annie Xu** & **Beijia Zhang**

COMS 6156:Topics in Software Engineering

Problem

- Traditional multimodal systems (audio, text, MIDI) assume all inputs are **equally reliable**.
- In real-world environments, inputs often become **noisy, corrupted, or incomplete** due to background noise, sensor failure, or transmission errors.
- Static fusion models that treat all modalities equally fail to adapt when one or more modalities degrade, leading to poor performance.

Solution: MuseMood

- **Dynamic Modality Weighting Module:**
  - Evaluates the **real-time quality** of each modality input.
  - **Adaptively adjusts** fusion weights based on modality reliability.
- **Dataset Preparation:**
  - Created clean and degraded versions of audio, lyrics, and MIDI.
  - Randomly degrade only one modality per sample to simulate real-world noise scenarios.
- **Robust Fusion:**
  - Trusts cleaner modalities more and minimizes the influence of degraded inputs during prediction.

Dataset

Audio

.mp3

Text

.txt

MIDIs

.mid

Sample Degraded Audio

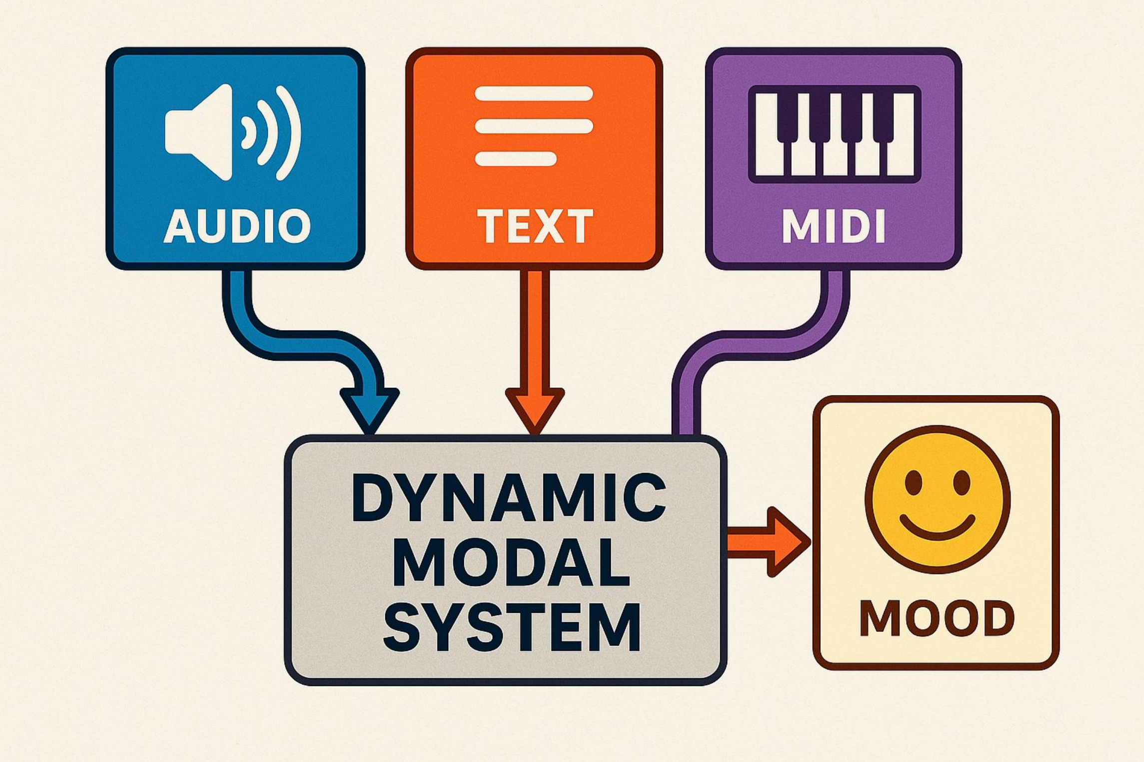
✅

❌

Model Used

- **Baseline Fusion Model**
  - Simply concatenates features from Audio, Lyrics, and MIDI.
  - Feeds them into a basic Feedforward Neural Network (FNN).
  - Treats all modalities **equally**, without knowing if data is noisy.
- **Dynamic Fusion Model**
  - Dynamically **reduces the weight** of degraded inputs during training and inference.
  - Similar neural network as baseline but with a **dynamic weighting mechanism** before fusion.

Dynamic Multi-Model



How We Assess Modality Quality

Audio (MP3):

- Measure features like Signal-to-Noise Ratio (SNR) or energy levels. Noisy or low-energy signals indicate degraded audio.

Text (Lyrics):

- Check for missing content, corrupted text, or unusually short/incomplete lyrics. High missing rates or blank sections suggest degraded text input.

MIDI (.mid):

- Analyze the structure — missing notes, abnormal timing, or incomplete tracks.
- Gaps or missing musical events signal degradation.

Baseline Model

Evaluation Metrics:

Accuracy : 0.3834

Precision: 0.4622

Recall : 0.3834

F1-Score : 0.3005

Dynamic Model

Evaluation Metrics:

Accuracy : 0.4819

Precision: 0.5178

Recall : 0.4819

F1-Score : 0.4769

Limitation

- **Small Dataset Size:**
  - Only 193 full samples were available after filtering. Limited data may affect generalization and stability of results.
- **Artificial Degradation Simulation:**
  - Modality degradation was synthetically generated and may not perfectly match real-world noise, limiting robustness validation.
- **Simple Model Architecture**
  - The basic MLP (Multi-Layer Perceptron) model may not capture complex cross-modal relationships as effectively as more advanced architectures like transformers or attention-based fusion.

Future Work

1. Expanding Dataset Size
2. Incorporating Real-World Degradation
3. Enhancing Model Complexity
4. Fine-Grained Emotion Prediction

Real-World Applications

- Emotion recognition from noisy music recordings
- Robust music video understanding and tagging
- Adaptive live performance systems
- Stronger music information retrieval under degraded input
- Smarter personalized music recommendation and mood detection

Thank You!