ELEC-E5510: Speech Recognition

Project report: Multimodal Emotion Recognition in Conversations

Cuong Nguyen (101559968), Raihan Gafur (101555441),
Daniel Campillejo Fernández-Calvillo (102591488)

2025-11-01

Contents

1 Introduction

Emotion recognition is a rapidly growing field that plays a pivotal role in advancing human-computer interactions. By enabling machines to interpret and respond to human emotions, it opens up possibilities for more natural and empathetic interfaces across diverse applications, such as healthcare, customer service, education, and entertainment [3, 1]. The challenge lies in accurately identifying emotional states, which are often expressed through a combination of verbal and non-verbal cues, including spoken words, vocal tone, facial expressions, and body language [3, 5].

While traditional emotion recognition systems typically rely on unimodal data, such as text or audio alone, they often fail to capture the complexity and nuances of human emotions [5, 1]. Multi-modal emotion recognition systems have emerged as a promising alternative, integrating multiple data sources to enhance accuracy and robustness [3, 5]. By combining information from text, audio, and visual modalities, these systems leverage the complementary nature of various emotional expressions to provide a more holistic understanding of emotional states [3, 1].

Despite advances in multi-modal recognition, several challenges persist. One major obstacle is the effective fusion of data from different modalities while preserving their individual characteristics and capturing their inter-modal relationships [3, 5]. Additionally, many datasets used to train these systems are curated in controlled environments, potentially limiting their generalizability to real-world interactions [3, 1].. Addressing these challenges requires innovative models and methodologies that are both computationally efficient and capable of extracting meaningful features from various data sources [5].

This project explores the development of a multi-modal emotion recognition system designed to overcome these limitations [3, 1].. Through the integration of advanced feature extraction techniques and hybrid fusion strategies, it aims to improve the accuracy and generalizability of emotion recognition models [3, 5]. By building on state-of-the-art research, this work seeks to contribute to the ongoing evolution of more intelligent and emotionally aware systems [5, 1].

2 Literature Study

Computers have become integral to daily life in areas such as healthcare, education, and entertainment. To interact effectively, machines must understand the emotional states of users. For example, in healthcare, recognizing emotions can help provide better support to patients, while in education, it enables personalized learning. Emotion recognition is thus vital for improving human-computer interaction. Research has explored various methods to recognize emotions using data sources like facial expressions, speech, text, and physiological signals. Among these, physiological signals are considered more objective and reliable [1]. Human emotions are broadly categorized into six fundamental types: happiness, sadness, fear, anger, disgust, and surprise, which form the basis for other derived emotions [5].

Unimodal emotion recognition systems utilize data from a single source but often fail to fully capture the complexity of emotions. For instance, relying solely on facial expressions may yield inaccurate results due to external influences like lighting or occlusion. To address these limitations, researchers have developed multimodal emotion recognition systems that combine multiple data sources. These systems outperform unimodal methods by providing a more comprehensive understanding of emotions [3, 1]. The efficiency and accuracy of deep learning models have made them particularly popular in this field, especially for analyzing high-level features in facial images and text [3].

There are three key multimodal fusion methods used in emotion recognition systems [3]:

- **Feature-level fusion**: Combines features from different modalities into a unified representation, allowing joint processing by the classifier.
- **Decision-level fusion**: Trains classifiers separately for each modality and merges their outputs using a fusion rule.
- **Hybrid fusion**: Combines aspects of feature-level and decision-level fusion to leverage their respective strengths.

Feature-level fusion is the most widely adopted approach in deep learning-based systems, as it integrates rich and complementary information from multiple modalities early in the processing pipeline, leading to higher accuracy by capturing subtle cross-modal emotional cues [3].

3 Methods

In this project, we focus on a multi-modal approach to emotion recognition, utilizing audio, text, and video data. Our primary aim is to integrate features from these modalities in a manner that preserves their unique characteristics while capturing the inter-modal relationships. To achieve this, we employ an **early fusion** approach, where features from all modalities are concatenated before being fed into the classifier. Below, we describe the key aspects of our methodology:

Feature Extraction The feature extraction process for each modality was designed to maximize the quality and relevance of input data:

- **Text Modality**: Features were derived using 300-dimensional FastText embeddings . These embeddings were further processed using a Convolutional Neural Network (CNN) to capture nuanced textual semantics and contextual information, building on advancements in natural language processing.
- **Audio Modality**: Audio features were extracted using the openSMILE toolkit , resulting in 6,373 initial features. These were reduced to 300 dimensions using L2-based feature selection, ensuring computational efficiency while retaining the most informative attributes for emotion detection.
- **Visual Modality**: For video data, 342 features were extracted using a DenseNet architecture pre-trained on the Facial Emotion Recognition Plus (FER+) corpus. These features represent critical visual cues such as facial expressions and micro-movements relevant to emotional states.

Early Fusion In this work, we utilize **early fusion** as the sole method for combining multi-modal features. Features from audio, text, and video modalities are concatenated into a unified representation and directly fed into the classifier. This approach ensures that the interactions between modalities are captured at the feature level while simplifying the architecture compared to decision-level fusion techniques.

Classifier We utilize the Bidirectional DialogueRNN with Attention [2] as the backbone of our classification system. This variant of DialogueRNN has demonstrated superior performance and was implemented as the baseline for our experiments. It is a speaker-aware recurrent neural network designed specifically for emotion recognition in multi-party conversations.

Dataset To train and evaluate this model we used MELD (Multimodal emotionless dataset) [4], this dataset is derived from over 1400 dialogues of the TV show Friends with more than 13.000 utterances, we used MELD because of its multimodal nature as it includes text, audio and visual data. The emotions we detect are: neutral, surprise, fear, sadness, joy, disgust and anger.

Evaluation The performance of our model is evaluated using weighted F1-scores and confusion matrices. We compare the results of our trimodal model with the baseline unimodal and bimodal models provided by MELD . Additionally, we explore the individual contributions of each modality by training unimodal classifiers for text, audio, and video separately.

4 Experiments

This section provides a detailed description of the experimental setup, including data loading, training, and testing procedures. The goal is to enable replication of our experiments by individuals familiar with the course material. A quick tutorial for training and testing is available here.

The baseline implementation lacks visual modality support and flexibility in handling multiple modality combinations. Our contributions address these gaps by incorporating visual features into the data loader and enhancing training and testing scripts for usability and transparency.

We conducted training and testing on the following modality combinations: audio, text, visual, text-audio, text-visual, audio-visual, text-audio-visual.

4.1 Data Loading

We utilized a pre-extracted pickle file containing MELD dataset features across three modalities and their alignment with video IDs, which was downloaded from this source.

To support diverse modality combinations, we modified the data loader's functions for item returning and batch collation to handle varying numbers of modalities and feature types. This enhancement ensures flexibility in experimentation.

4.2 Training

Our training script extends the baseline implementation by adding support for multiple modality combinations and providing detailed metrics for each epoch. The best performance metrics, achieved at the epoch with minimal error, are reported. To ensure reproducibility, the random seed is fixed at 1234.

An example command to train the model is shown in Listing 1

4.3 Testing

We implemented a custom testing script for model evaluation. The script uses a saved model to evaluate performance on the test set. The final performance may differ from training due to metric calculations on different epochs.

To test the model, execute the command in Listing 2

5 Results

The experiments evaluated the performance of emotion recognition models using three different modality combinations: unimodal, bimodal, and trimodal approaches. The key metrics for evaluation were weighted F1-score and accuracy. The following summarizes the results for each modality combination:

5.1 Summary Table

Type	Modality Combination	Weighted F1-score	Accuracy (%)
Unimodal	Audio	43.55	48.85
	Visual	35.35	44.75
	Text	57.52	59.77
Bimodal	Audio-Visual	45.25	48.24
	Text-Audio	57.61	59.39
	Text-Visual	57.51	60.15
Trimodal	Text-Audio-Visual	57.88	60.61

Table 1: Performance summary across different modality combinations.

The results summarized in Table 1 demonstrate that integrating multiple modalities improves emotion recognition performance. While unimodal models showed reasonable accuracy, they struggled to capture the full complexity of emotional cues. Bimodal combinations, particularly those involving text, provided marginal improvements over their unimodal counterparts. The trimodal fusion *(Text-Audio-Visual)* achieved the highest accuracy and F1-score, reflecting the benefits of leveraging complementary information across all three modalities.

5.2 Confusion Matrices and Class Performance

The confusion matrices reveal that dominant classes like *Neutral* (Class 0) and *Happy* (Class 1) consistently achieved high recall across all modalities due to their abundance in the dataset. In contrast, underrepresented emotions such as *Fear* (Class 2) and *Disgust* (Class 5) exhibited near-zero recall and were frequently misclassified as *Neutral* or *Sadness*. Misclassifications also occurred between similar emotions, such as *Anger* (Class 3) and *Sadness* (Class 4), particularly in unimodal and bimodal setups. While trimodal fusion reduced these ambiguities by leveraging complementary information, categories like *Surprise* (Class 6) still faced confusion, often being misclassified as *Neutral*.

The model's tendency to default to the *Neutral* class when uncertain highlights the impact of dataset imbalance on predictions. Addressing

this issue through techniques like class weighting, oversampling, or advanced fusion methods, such as attention mechanisms, could enhance performance on rare and overlapping emotion categories.

The seven confusion figures are presented in the appendices 1, illustrating the classification performance across different modalities in the specified order.

6 Conclusion

This study explored the effectiveness of multimodal emotion recognition by combining text, audio, and visual data. The results highlight the value of multimodal approaches, with the trimodal configuration (text-audio-visual) achieving the highest performance. However, the improvements over text-only and bimodal models were marginal, suggesting limitations in the early fusion strategy used. Advanced fusion techniques like attention mechanisms may better exploit inter-modal interactions.

The text modality proved to be the most reliable, both alone and in combination with other modalities, reflecting its richness in conveying emotional cues. In contrast, underrepresented emotions like *Fear* and *Disgust* showed near-zero recall and precision due to dataset imbalance. These findings emphasize the need for class balancing techniques, such as weighted loss functions or data augmentation, to improve the recognition of rare emotions.

Misclassifications between similar emotions, such as *Anger* and *Sadness*, further underscore the limitations of the current feature extraction methods. The model's bias toward the dominant *Neutral* class, influenced by dataset imbalance, also points to the need for more balanced datasets and context-aware modeling.

In summary, while multimodal systems show promise, challenges remain in inter-modal fusion, class imbalance, and generalizability. Future work should focus on improving fusion strategies, addressing dataset limitations, and leveraging advanced architectures to enhance the robustness and applicability of emotion recognition models.

7 Division of labor

- Cuong Nguyen:
  - Wrote experiment code.
  - Wrote ‘Literature Study’ and ‘Experiments’ sections.
- Raihan Gafur:
  - Wrote classification matrices code.
  - Wrote ‘Results’ and ‘Conclusion’ sections.
- Daniel Campillejo Fernández-Calvillo:
  - Wrote unimodal implementation (discarded in the end).
  - Wrote ’Introduction’ and ’Methods’ sections.

8 Acknowledgments

We extend our heartfelt thanks to Guangpu Huang for proposing this fascinating research topic and for providing invaluable guidance in introducing us to the field of multimodal emotion recognition and the MELD dataset. In addition, his informative feedback has helped us get the project on the right track.

References

[1] K. Ezzameli and H. Mahersia. "Emotion recognition from unimodal to multimodal analysis: A review". In: *Information Fusion* 99 (2023), p. 101847. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101847. URL: https://www.sciencedirect.com/science/article/pii/S156625352300163X.

[2] Navonil Majumder et al. "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations". In: *CoRR* abs/1811.00405 (2018). arXiv: 1811.00405. URL: http://arxiv.org/abs/1811.00405.

[3] Bei Pan et al. "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods". In: *Neurocomputing* 561 (2023), p. 126866. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126866. URL: https://www.sciencedirect.com/science/article/pii/S092523122300989X.

[4] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *CoRR* abs/1810.02508 (2018). arXiv: 1810.02508. URL: http://arxiv.org/abs/1810.02508.

[5] Jianhua Zhang et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59 (2020), pp. 103–126. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.01.011. URL: https://www.sciencedirect.com/science/article/pii/S1566253519302532.

Appendices

A Code

All code is publicly accessible at `https://github.com/ancuongnguyen07/Multimodal-ERC/tree/master/code`

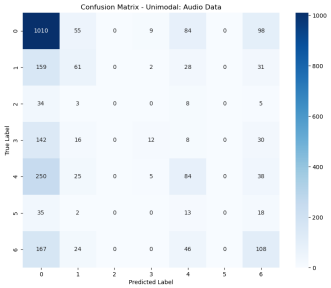
Listing 1: Running the training script.

```
python train_MELD.py --features-type text_audio_visual \
    --data-path data/MELD_features_raw1.pkl --output-dir models/
```

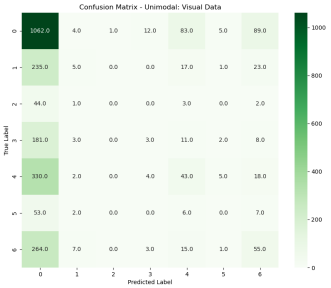
Listing 2: Running the testing script.

```
python test_MELD.py --model-path ../models/
    text_audio_visual_BiDi_Att.pth \
    --features-type text_audio_visual
```

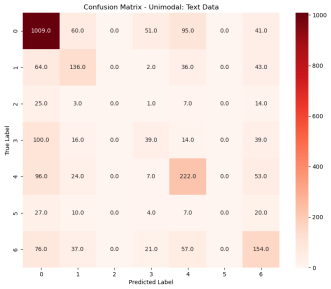
B Confusion Matrices



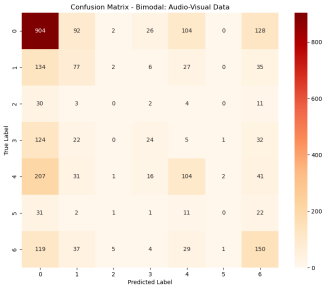
(a) Unimodal: Audio Data



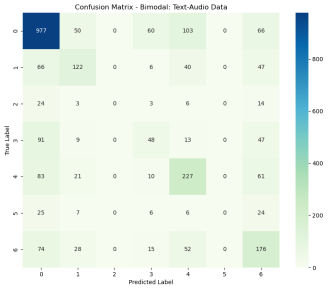
(b) Unimodal: Visual Data



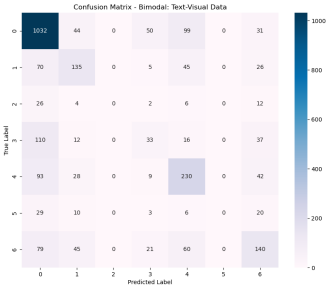
(c) Unimodal: Text Data



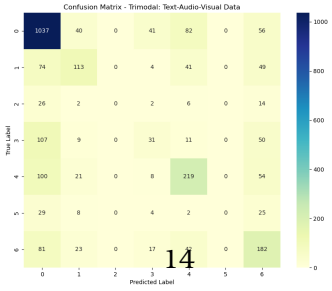
(d) Bimodal: Audio-Visual Data



(e) Bimodal: Text-Audio Data



(f) Bimodal: Text-Visual Data



(g) Trimodal: Text-Audio-Visual Data

Figure 1: Confusion Matrices Across Modalities