Multimodal Emotion Recognition Using Text-Video

Dang Tran Tan Luc, Phung Khanh Duy

July 11, 2025

Abstract

Multimodal sentiment analysis (MSA) aims to use a variety of sensors to obtain and process information to predict the intensity and polarity of human emotions. The main challenges faced by current multi-modal sentiment analysis include: how the model extracts emotional information in a single modality and realizes the complementary transmission of multimodal information; how to output relatively stable predictions even when the sentiment embodied in a single modality is inconsistent with the multi-modal label; how can the model ensure high accuracy when a single modal information is incomplete or the feature extraction performance not good. Traditional methods do not take into account the interaction of unimodal contextual information and multimodal information. They also ignore the independence and correlation of different modalities, which perform poorly when multimodal sentiment representations are asymmetric. To address these issues, we will proposes an unimodal feature extraction network to extract unimodal features with stronger representation capabilities; then introduces multi-task fusion network (MTFN) to improve the correlation and fusion effect between multiple modalities. Multilayer feature extraction, attention mechanisms and Transformer are used in the model to mine potential relationships between features. This project performs multimodal sentiment analysis using the CMU-MOSEI dataset, using transformer based models with early fusion to integrate text and visual modalities. We will employ BERT-based encoders for each modality, extracting embeddings that are concatenated before classification

1 Introduction

Effective human communication hinges on the fundamental skill of emotional awareness, enabling individuals to comprehend and respond to the feelings of others. Conventional emotion recognition methods [1] often rely on data from a single source, such as speech patterns or facial expressions. However, human emotional communication is inherently multimodal, involving a complex interplay of verbal and non-verbal cues. To address this complexity, recent research has increasingly focused on multimodal emotion recognition tasks that combine audio and video sources. By leveraging both verbal intonations and facial expressions [2], these approaches aim to provide a deeper and more nuanced understanding of an individual's state

Consider the example: "I'm really happy for you!" Determining whether genuine happiness is being expressed based solely on the tone of voice can be challenging. Visual cues, such as a sincere smile, wrinkled eyes, and an overall positive demeanor, offer additional insights. However, relying solely on visual data has its drawbacks, as a smile can be faked to mask true feelings. Therefore, combining audio and visual elements is essential to capture a more accurate emotional context [3], [4].

help library, or head to our plans page to choose your plan.

2 Related Works

2.1 MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos [5]

This paper focuses on the problem that while sentiment analysis has been successful for text, This is an understudied research question for video and other multimedia content. A major obstacle to this research is the lack of appropriate datasets, methodologies, baselines, and statistical analysis of how information from different methodological sources relate to each other.

This paper introduces to the scientific community the first opinion-level annotation dataset for sentiment and subjectivity analysis in online videos, called the Multimodal Opinionlevel Sentiment Intensity dataset (MOSI). The dataset is rigorously annotated with subjective labels, sentiment intensity, frame-by-frame, opinion-by-opinion, and millisecond-annotated audio features. They present baselines for future research in this direction as well as a novel multimodal fusion approach that jointly models spoken words and visual gestures.

2.2 Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention [6]

The AVT-CA model presents a robust solution for multimodal emotion recognition by addressing synchronization, feature extraction, and fusion challenges. Its innovative use of channel and spatial attention, local feature extraction, transformer fusion, and cross-attention mechanisms enables precise and reliable emotion recognition. The model's superior performance on CMU-MOSEI, RAVDESS, and CREMA-D datasets highlights its potential for applications in human-computer interaction, entertainment, and mental health monitoring. Future work could extend this approach to other multimodal tasks or refine attention mechanisms for even better performance.

2.3 Multimodal Language Analysis in the Wild: CMU-MOSEIDataset and Interpretable Dynamic Fusion Graph [7]

The paper introduces the CMU-MOSEI dataset and the Dynamic Fusion Graph (DFG) model, two significant contributions to the study of multimodal language. CMU-MOSEI provides a robust platform for sentiment analysis and emotion recognition with its large scale and diversity in speakers and topics. The DFG, when integrated into Graph-MFN, not only achieves high performance but also provides deep insights into how modalities (language, visual, and acoustic) interact in human communication through efficacy coefficients. The findings show that DFG learns natural priors of communication, such as prioritizing language-acoustic fusion and selectively utilizing the visual modality. This approach opens new research directions for multimodal language analysis in NLP and multimodal machine learning

3 Problem Statement

3.1 Problem

The rapid advancement of human-computer interaction and affective computing has underscored the importance of understanding human emotions through multimodal data. Emotion recognition from short video clips, which combine visual, auditory, and textual cues, presents a complex yet critical challenge for applications in areas such as mental health monitoring, virtual assistants, and social robotics. The CMU-MOSEI dataset, a comprehensive multimodal dataset, provides a rich source of annotated video clips featuring human emotions expressed through facial expressions, speech, and language. This project aims to develop a machine learning model to accurately recognize emotions from short video clips using the CMU-MOSEI dataset, addressing the challenge of integrating and processing multimodal inputs to predict emotional states.

Inputs

The model takes as input multimodal data extracted from short video clips in the CMU-MOSEI dataset. Specifically, the inputs consist of:

Visual Features: Video frames capturing facial expressions and body gestures, typically processed as sequences of images or extracted features (e.g., facial landmarks or embeddings from pre-trained models).

Audio Features: Speech signals, including raw audio waveforms or derived features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and tone.

Textual Features: Transcribed spoken text or subtitles, often represented as word embeddings (e.g., GloVe or BERT embeddings) to capture semantic content.

Outputs

The output of the model is a predicted emotion label for each video clip, corresponding to one or more discrete emotional categories (e.g., happiness, sadness, anger, surprise, disgust, fear) or continuous emotional dimensions (e.g., valence, arousal, and dominance). The model may also provide confidence scores or probabilities associated with each predicted emotion, enabling a nuanced interpretation of emotional states.

3.2 Algorithm

1. **Feature Extraction with BERT [2]:** The use of a pre-trained BERT model (BertModel from bert-base-uncased) for text feature extraction is part of the supervised pipeline. While BERT itself was pre-trained in an unsupervised manner (using masked language modeling), in this context, it is fine-tuned or used to extract features for a supervised task (sentiment prediction), as the extracted embeddings are paired with sentiment labels for downstream classification.

    - **Implementation**: The BertTokenizer tokenizers input text into subword tokens, padding or truncating to a fixed length (TEXT_MAX_LENGTH = 128) The BertModel (pre-trained bert-base-uncased) extracts embeddings from tokenized text. Specifically, the token's output (first token in the sequence) is used as a 768-dimensional representation (TEXT-_EMBEDDING_DIM) of the text. The model is set to evaluation mode (eval()) and run with torch.no_grad() to avoid gradient computation, optimizing for feature extraction.
    - **Why used:** BERT provides robust, context-aware embeddings that capture semantic and emotional nuances in text, critical for sentiment analysis.

2. **Feature Extraction for Audio (COVAREP [4] Features):** Audio features from the CMU-MOSEI dataset (COVAREP features) are time-series data representing acoustic properties like pitch and energy. To obtain a fixed-size representation, the mean of these features is computed across the time dimension.

    - **Implementation:** The extract_audio_features method computes the mean of audio features along the time axis. Handles cases where the feature vector size does not match AUDIO_FEATURE_SIZE (40) by truncating or padding with zeros. Replaces NaN values with zeros to ensure numerical stability.
    - **Why used:** Mean aggregation simplifies variable-length time-series data into a fixed-size vector, making it compatible with downstream models while retaining key acoustic characteristics.

3. **Feature Extraction for Visual (OpenFace2 [3] Features):** Visual features from OpenFace2 in CMU-MOSEI include facial landmarks and expressions. Similar to audio, mean aggregation is applied to reduce time-series data to a fixed-size vector.

    - **Implementation:** The extract_visual_features method computes the mean of visual features along the time axis.
    - **Why used:** Mean aggregation ensures a consistent representation of visual features, capturing average facial expressions relevant to sentiment.

4. **Attention Mechanism [1]:** The attention mechanism is the core innovation of the Transformer, enabling it to weigh the importance of different parts of the input data when making predictions. In the context of the code, attention is used within the BERT model for text processing and likely in the multimodal Transformer for cross-modal fusion.

    Attention allows the model to focus on relevant parts of the input sequence when processing each element, rather than treating all parts equally. It computes a weighted sum of input features, where weights reflect the relevance of each input to the current context. The primary attention mechanism used in Transformers is scaled dot-product attention. Here's how it works:

    - **Inputs:**
        - **Query (Q)**: Represents the current token or feature being processed.

- **Key (K)**: Represents all tokens/features in the sequence to compare against.
- **Value (V)**: Contains the actual information to be weighted.

- **Linear Transformations:** These are derived from the input embeddings via linear transformations:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

where X is the input sequence and W_Q, W_K, W_V are learnable weight matrices.

- **Attention scores:**
  - Compute the similarity between the query and each key using a dot product: Score = QK^T.
  - Scale the scores by the square root of the key dimension ($\sqrt{d_k}$) to prevent large values: Scaled Score = $\frac{QK^T}{\sqrt{d_k}}$.
  - Apply a softmax to obtain attention weights: Attention Weights = softmax $\left(\frac{QK^T}{\sqrt{d_k}}\right)$.

- **Multi-Head Attention:**
  - Instead of computing attention once, multi-head attention splits the input into multiple subspaces (heads), applies scaled dot-product attention to each, and concatenates the results.
  - **Formula:** MultiHead(Q, K, V) = Concat$(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O$, where head_i = Attention$(QW_{Q_i}, KW_{K_i}, VW_{V_i})$, and W_O is a projection matrix.

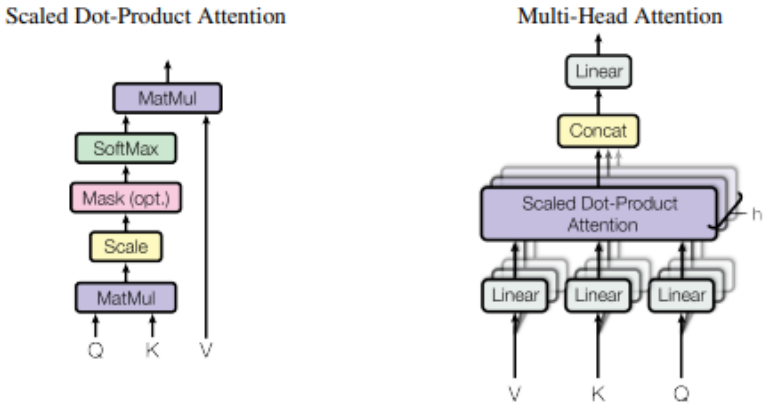

Figure 1: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

5. **Transformers Fusion:** The initial description mentions a Transformer-based model with early fusion to integrate text, audio, and visual modalities. While not implemented in the provided code, this is a key algorithm for the overall MSA pipeline. Once the data streams are processed, we merge them using a technique called early fusion. In early fusion, the features from each modality are combined before they are passed to the classifier. In this project, we extract fixed-length embeddings from each modality using pre-trained BERT-based encoders, we concatenate these vectors to form single joint feature vector:

$$H_{\text{fused}} = [H_{\text{text}}; H_{\text{visual}}]$$

Why used: Combines modalities early (via concatenation) and uses Transformer layers to model cross-modal interactions, addressing challenges like inconsistent or incomplete modalities.

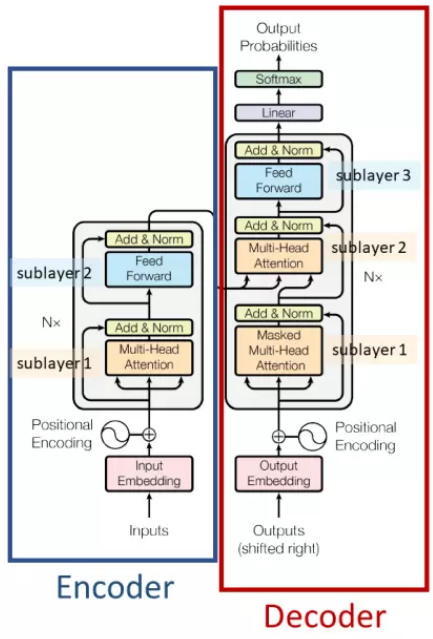

Figure 2: Transformers Architectural.

4 Experimental Evaluation

4.1 Dataset

For the project, we used the CMU-MOSEI (Multimodal Opinion Sentiment and Emotion intensity) dataset, a benchmark dataset created and maintained by Carnegie Mellon University. It is one of the largest benchmark dataset for multimodal sentiment analysis. It contrains rich data aligned across three modalities-textual, visual and acoustic.

The CMU-MOSEI dataset contains over 23,000 annotated sentence-level video segments from about 1,000 Youtube speakers. These segments are chosen randomly from various topics and monologue videos. All the data is annotated for both sentiment palarity and emotion intensity, making it a great choice for multimodel learning. Table 1 provides the dataset statistics.

Feature	Name
Total number of sentences	23453
Total number of videos	3228
Total number of distinct speakers	1000
Total number of disstince topics	250
Average number of sentences in a video	7.3
Average length of sentences in seconds	7.28
Total number of words in sentences	447143
Total of unique words in sentences	23026
Total number of words appearing at least 10 times	3413
Total number of words appearingat least 20 times	1971
Total number of words appearing at least 50 times	888

Table 1: CMU-MOSEI dataset statistics.

Figure 1 shows the diversity of topics of videos in CMU-MOSEI, displayed as a word cloud. Larger words indicate more videos from that topic.

Data features: The textual features in the dataset were extracted from the manual transcriptions using the Glove word embeddings. Both words and audio were aligned using the P2FA forced alignment model. The visual and acoustic modalities are aligned using interpolation.

Figure 3: Word cloud visualizing topic diversity in CMU-MOSEI

For the creation of visual features, the facial expressions were extracted from the video frames using MTCNN followed by the Facial Action Coding System (FACS). Then, the static faces were classified into six basic emotions: anger, disgust, fear, happiness, sadness and surprise using Emotient FACET. The data contains a set of 68 facial landmarks, 20 facial shape parameters, Histogram of Gradient (HoG) features, head pose, head orientation and eye gaze extracted using MultiComp OpenFace. Finall, the extracted face embeddings using DeepFace, FaceNet and SphereFace.

The acoustic data was extracted using COVAREP software. The extracted features include 12 Mel-frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients, all of which are related to emotions and tone of speech.

The sentiment labels in the dataset range from -3 to +3 (-3, -2, -1, 0, +1, +2, +3), where -3 means strong nagative sentiment, 0 means neutral, and +3 means strong positive sentiment.

The datase is already preprocessed and aligned across all three modalities. It is available for download through CMU Multimodal Data SDK at https://github.com/CMU-MultiComp-Lab/CMU-MultimodalSDK

Data Visualization

Figure 4: Correlation between Features and Sentiment Label: Most features exhibit weak or negligible correlations, fluctuating around zero, indicating limited linear dependency. A few features show stronger positive or negative correlations (approaching ±0.3), suggesting they may carry more discriminative information for sentiment prediction.

Figure 5: Sentiment Score Distribution: The histogram displays the distribution of sentiment scores on the test set. Most scores are concentrated between 0.0 and 0.3, indicating that the majority of samples express negative to mildly neutral sentiment. The distribution is right-skewed, with relatively few samples exhibiting high sentiment scores, corresponding to strongly positive emotions.

Figure 6: Correlation Matrix - Visual Features: The correlation matrix above illustrates the relationships between visual features. Red cells indicate high positive correlations (close to +1), while blue cells represent strong negative correlations (close to -1). The results show that the features from column 12 onwards are highly correlated with each other, suggesting potential redundancy in this feature set.

Figure 7: Correlation Matrix - Text Features: The correlation matrix above shows the relationships between text features. Most feature pairs have low correlation values (close to zero), as indicated by the dominance of light colors in the matrix. This suggests that the text feature set is diverse and does not contain strong redundancy among its components.

Figure 8: Correlation Matrix - Audio Features: The correlation matrix above presents the relationships between audio features. Some feature pairs show moderate correlations, as seen in the scattered red and blue regions. Overall, the feature set exhibits a mix of correlated and uncorrelated

components, indicating partial redundancy but also diverse information among features.

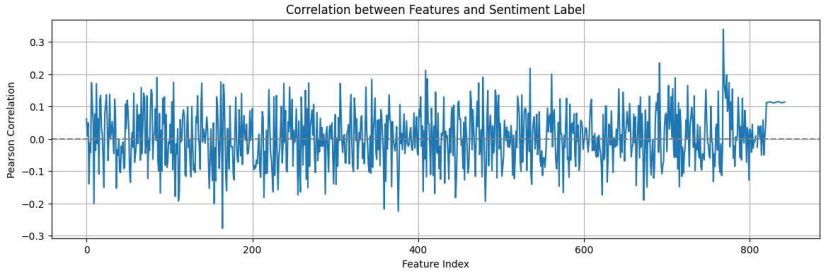

Figure 4: Correlation between Features and Sentiment Label

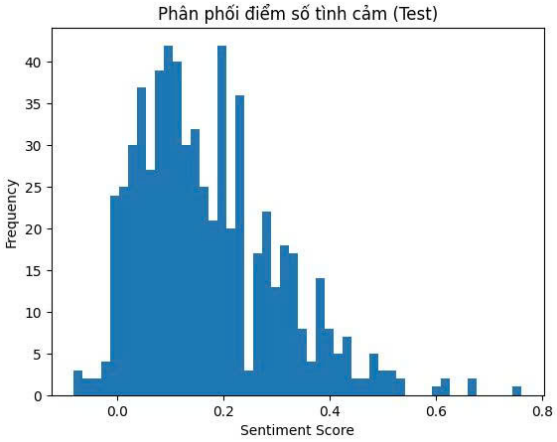

Figure 5: Sentiment Score Distribution

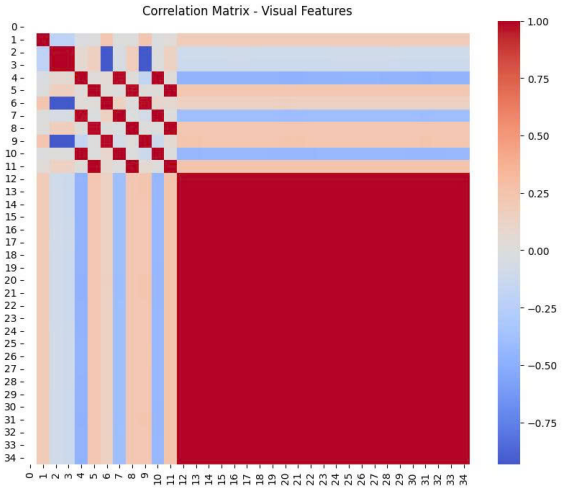

Figure 6: Correlation Matrix - Visual Features

Data Issues: Several challenges in the CMU-MOSEI dataset necessitate preprocessing:

- Missing Data: Some clips lack complete modality data (e.g., silent audio or missing frames), requiring imputation or exclusion.
- Class Imbalance: Underrepresented emotions (e.g., surprise, disgust) may bias model performance toward dominant classes.
- Variable Sequence Lengths: Inconsistent numbers of frames, audio segments, or tokens across clips complicate input alignment.
- Noise and Artifacts: Low-resolution frames, background noise in audio and transcription errors in text introduce noise.
- Dimensionality Mismatch: Features from different modalities (e.g., ResNet vs BERT) have varying dimensions, necessitating alignment.

Rational for Data Preprocessing: The identified issues provide the basis for applying preprocessing techniques:

- Handling Missing Data: Imputation or exclution ensures rebust multimodal imputs.
- Addressing Class Imbalance: Techniques like weighted loss or oversampling improve model fairness.
- Sequence Alignment: Padding or truncation standardizes input lengths for Transformer processing.

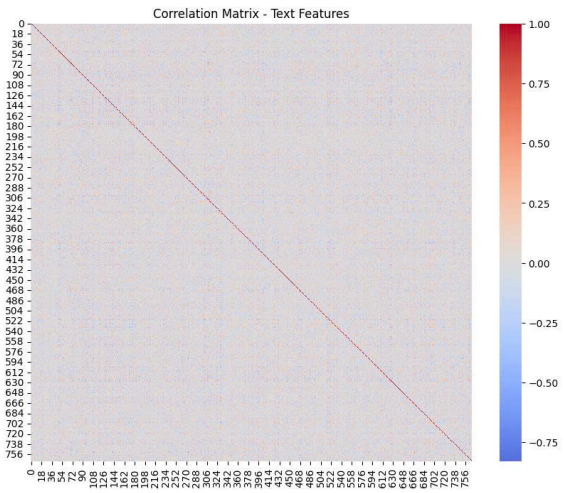

Figure 7: Correlation Matrix - Text Features

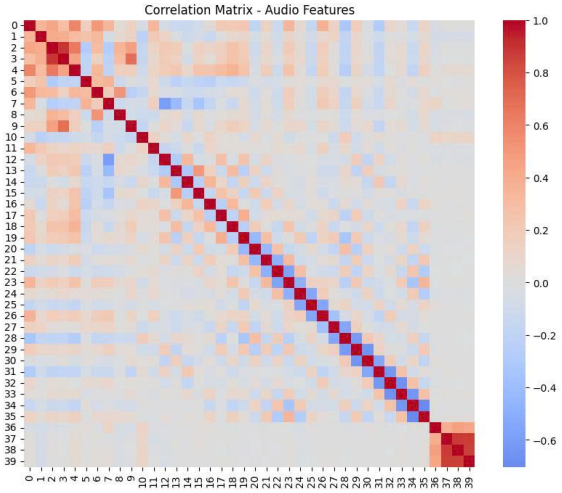

Figure 8: Correlation Matrix - Audio Features

- Noise Reduction: Filtering and normalization enhance data quanlity.
- Feature Alignment: Dimensinality reduction or projecttion ensures compatibility across modalities.

These preprocessing steps are essential to mitigate biases, improve model convergence, and enable effective training on the CMU-MOSEI dataset.

The dataset was divided into three subsets: a training set, a validation set, and a testing set. The training set was used to optimize model parameters, the validation set was employed to tune hyperparameters and prevent overfitting, and the testing set was reserved for evaluating the final performance of the model. The data split was performed to ensure that each subset maintained a similar distribution of samples across relevant classes and conditions. The dataset was split into 70% for training, 10% for validation, and 20% for testing.

4.2 Results

To assess the effectiveness of our proposed multi-model sentiment analysis model, we evaluated its performance on the training, validation, and test sets using a comprehensive set of metrics. These include Mean Absolute Error (MEA), Accuracy, F1-Score, Binary Accuracy, and Binary F1 Score. Additionally, the training dynamics were monitored over 5 epochs to stable convergence and generalization

These metrics reflect a high degree of fidelity in both fine-grained sentiment classification (across 8 classes from -3 to +3) and binary sentiment polarity section (positive and negative)

Model	MAE	Corr	Binary Acc	AccuracyBinary F1
LateFusion	0.10325101763010025	0.38286336326128706	0.94140625	0.9698189134808853
TransformerFusion	0.08131558448076248	0.5701004013611962	0.96484375	0.9821073558648111

Figure 9: Validation metrics

The performance of the proposed LateFusion and TransformerFusion models was evaluated on the CMU-MOSEI dataset using multiple metrics, including Mean Absolute Error (MAE), Correlation (Corr), Binary Accuracy (Binary Acc), and AccuracyBinary F1 score. The results are summarized as follows:

Mean Absolute Error (MAE): TransformerFusion achieved a lower MAE of 0.081316 compared to 0.103251 for LateFusion, indicating a 21% reduction in prediction error for continuous emotion dimensions. This suggests improved accuracy in regression tasks.

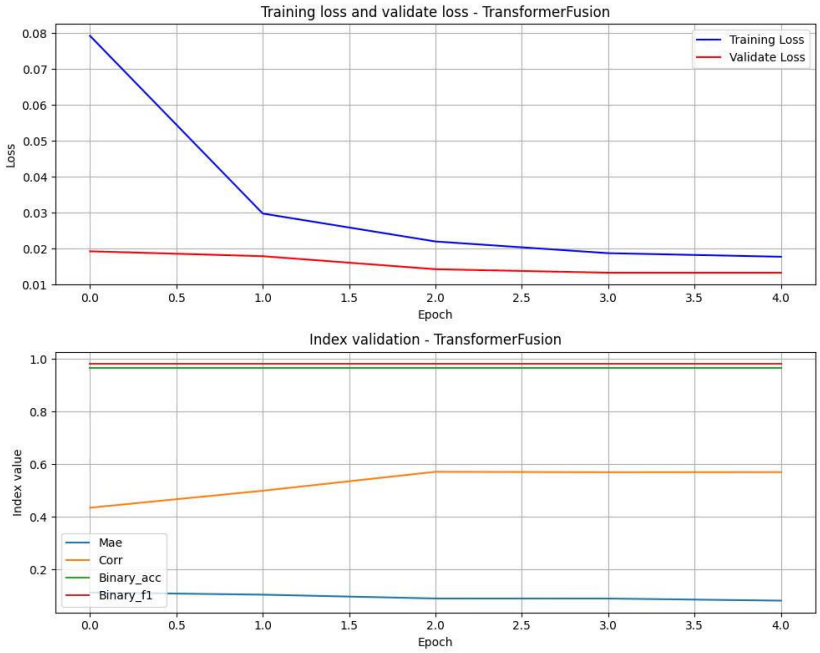

Figure 10: Training loss and validate loss

Correlation (Corr): TransformerFusion demonstrated a higher correlation coefficient of 0.5701 versus 0.3828 for LateFusion, reflecting a 51% improvement in capturing the linear relationship between predicted and ground-truth values.

Binary Accuracy (Binary Acc): TransformerFusion outperformed LateFusion with a Binary Acc of 0.9648 (96.48%) compared to 0.9414 (94.14%), a relative increase of 2.34%, highlighting better performance in binary emotion classification.

AccuracyBinary F1: TransformerFusion recorded a higher F1 score of 0.9821 versus 0.9698 for LateFusion, an improvement of 1.25%, indicating a better balance between precision and recall in binary classification tasks.

Overall, TransformerFusion consistently outperformed LateFusion across all evaluated metrics, with improvements ranging from 1.25% (F1) to 51% (Corr). This superior performance can be attributed to the Transformer's ability to model complex cross-modal interactions through self-attention mechanisms, in contrast to the late fusion approach, which combines modality-specific features at a later stage and may overlook intricate multimodal relationships. These findings underscore the effectiveness of the Transformer-based architecture for multimodal emotion recognition on the CMU-MOSEI dataset.

5 Conclusion

In this project, we presented a transformer-based multimodal sentiment nalysis model that uses an early fusion of textual, acoustic, and visual features to effectively predict sentiment on the CMU-MOSEI dataset. By combining modality-specific transformer-based encoders with a classification head enhanced by Layer Normalization, Dropout, and ReLU activation, the model achieved strong generalization and high accuracy across both binary and 8-class sentiment prediction tasks.

Our approach demonstrated consistent improvements across validation and test sets, achieving 0.9648 accuracy and 0.9821 F1-score on the 8-class sentiment test set, alongside a low MAE of 0.081. These results highlight the model's ability to not only learn cross-modal relationships effectively but also avoid overfitting through regularization techniques and efficient model design.

The model's performance shows the importance of early fusion strategies when combined with pretrained BERT encoders and proper regularization for multimodal tasks. Future work may explore the integration of attention-based fusion mechanisms or a comparison between early fusion and late fusion techniques to further enhance interpretability and performance.

article

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention is All You Need*. In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In NAACL-HLT, 2019.

[3] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. *OpenFace 2.0: Facial Behavior Analysis Toolkit*. In IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2018.

[4] Philippe Le Digabel Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stylianos M. Georgiou. *COVAREP – A Collaborative Voice Analysis Repository for Speech Technologies*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

[5] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. *MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016.

[6] Xuefeng Li, Xiaofeng Mao, Yanyan Zou, Linlin Shen, and Ling Shao. *Multimodal Emotion Recognition using Audio-Video Transformer Fusion with Cross Attention*. IEEE Transactions on Multimedia, 2023.

[7] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. *Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph*. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018.

Link Notebook Link Colab of our code (click here)