Two-Stage Emotion Detection from Multimodal Data

A Project Report

Presented to
The Faculty of the Department of Computer Science
San Jose State University

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science

By
Xiangyi Li
Spring 2024

The Designated Project Committee Approves the Project Titled

Two-Stage Emotion Detection from Multimodal Data

by

Xiangyi Li

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

Spring 2024

Prof. Faranak Abri	Department of Computer Science
Prof. Fabio Di Troia	Department of Computer Science
Ms. Shuyi Wang	International Monetary Fund

Abstract

Emotion detection plays a crucial role in human-computer interaction, enabling machines to recognize and respond appropriately to human emotional states. This project explores a two-stage approach to emotion detection using multimodal data, specifically analyzing both textual content and audio features. We investigate different model architectures including transformer-based language models like BERT, RoBERTa, and DeBERTa, alongside various fusion techniques for combining modalities. Using the IEMOCAP dataset in both its complete and filtered versions, we evaluate the effectiveness of single-modality versus multimodal approaches. Our findings demonstrate that transformer-based models, particularly RoBERTa, achieve the highest accuracy when processing textual data alone, while hybrid and late fusion methods yield superior performance when combining text with audio features, particularly MFCC and spectrogram representations. The optimal configuration combines the RoBERTa model with MFCC features using hybrid fusion, achieving an impressive validation accuracy of 91.74%. This research contributes valuable insights into the relative importance of different modalities and fusion strategies for emotion recognition systems, with potential applications in affective computing, mental health monitoring, and more intuitive human-machine interfaces.

Keywords: Emotion Detection, Natural Language Processing, Multimodal Analysis, Audio Processing, Transformer Models, Fusion Techniques

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Faranak Abri, for her invaluable guidance and support throughout this project. Her insights and feedback have helped me navigate challenges and achieve my goals. Thank you for your mentorship and encouragement.

I would like to extend my gratitude to all members of my defense committee, Professor Fabio Di Troia, and Ms. Shuyi Wang.

I would like to extend my heartfelt thanks to my family and friends for their unwavering support and encouragement throughout my academic journey and in all aspects of my life.

Contents

List of Figures

List of Tables

1 Introduction

Emotion recognition plays a fundamental role in human communication, allowing us to understand others' feelings, intentions, and needs. As artificial intelligence systems become increasingly integrated into our daily lives, the ability for machines to recognize and respond appropriately to human emotions has become crucial for meaningful human-computer interaction. This capability, often referred to as affective computing, has applications ranging from healthcare and education to entertainment and customer service.

Emotions are expressed through multiple channels, including facial expressions, voice modulations, body language, and verbal content. While humans naturally process these cues simultaneously to gauge emotional states, developing computational systems that can effectively interpret and integrate these diverse signals remains challenging. Traditional approaches often focus on a single modality, such as analyzing facial expressions or processing textual content, which limits their robustness across different contexts.

The challenge in emotion detection stems from several factors. First, emotions are inherently subjective and exist on a spectrum rather than as discrete categories. Second, emotional expressions vary considerably across individuals, cultures, and contexts. Third, different modalities may convey conflicting emotional information, requiring sophisticated fusion techniques to resolve inconsistencies. Furthermore, environmental factors like background noise or lighting conditions can significantly impact the quality of input signals.

To address these challenges, this project explores a two-stage approach to emotion detection using multimodal data. The first stage involves processing individual modalities—specifically text and audio features—through specialized models tailored to each data type. The second stage employs various fusion techniques to combine these modality-specific representations, leveraging their complementary strengths while mitigating their individual weaknesses.

Our research makes several key contributions:

1. Comparative Analysis of Language Models: We evaluate the performance of various transformer-based models, including BERT, RoBERTa, XLNet, ALBERT, ELECTRA, and DeBERTa, for textual emotion recognition.

2. Audio Feature Exploration: We investigate the effectiveness of different audio representations, including Mel-Frequency Cepstral Coefficients (MFCC), spectrograms, prosodic features, and wav2vec embeddings.

3. Fusion Strategy Assessment: We systematically compare early, late, hybrid, and attention-based fusion approaches for integrating textual and audio modalities.

4. Benchmark Dataset Evaluation: We conduct experiments on both the complete and filtered versions of the IEMOCAP dataset, providing insights into the impact of data preprocessing on model performance.

5. Modal Cloud Infrastructure: We implement a scalable and reproducible experimental framework using Modal's cloud infrastructure, enabling efficient parallel execution of numerous experiments.

The structure of this report is as follows: Section **??** reviews relevant literature on emotion recognition approaches. Section 3 describes our methodology, including model architectures and fusion strategies. Section 4 details the experimental setup, covering dataset preparation, preprocessing techniques, and evaluation metrics. Section 5 presents our findings, while Section 6 discusses their implications. Finally, Section 7 concludes the report and suggests directions for future research.

2 Related Work

2.1 Early Emotion-Recognition Approaches (pre-2012)

The first wave of emotion-recognition (ER) research was dominated by lexicon and rule-based techniques such as WordNet-Affect [1] and the NRC Emotion Lexicon [2]. Classical machine-learning models—Naïve Bayes, logistic regression and SVM—were trained on bag-of-words, TF–IDF or LIWC counts for text ER, and on low-level descriptors for speech or facial data [3]. These systems rarely exceeded ≈60 % accuracy on early benchmarks and were brittle to contextual nuance.

2.2 Deep-Learning Era (2013–2017)

Convolutional and recurrent networks rapidly eclipsed classical models. CNNs and (Bi-)LSTMs learned richer semantic and temporal features for text, audio and vision. In speech ER, CNN-style spectrogram encoders combined with LSTM temporal heads pushed performance above 90 % on Emo-DB and IEMOCAP [4]. Hybrid CNN-LSTM architectures similarly boosted textual ER [5]. Multimodal early-fusion CNNs that concatenate openSMILE prosody, word embeddings and 3-D CNN facial features achieved ≈80 % on CMU-MOSI [6]. Although effective, these networks struggled with asynchronous modalities and long-range context.

2.3 Transformer-Based Models (2018–2025)

Contextualised language models revolutionised unimodal ER. Fine-tuning BERT, RoBERTa or DeBERTa yields state-of-the-art textual accuracies on many datasets [7]. Cross-modal transformers extend self-attention to heterogeneous streams. MulT [8] attends across misaligned audio, visual and linguistic tokens, attaining 84.8 % unweighted accuracy (UA) on the six-class IEMOCAP task. Progressive-modality reinforcement [9] iteratively re-weights modalities, while TransModality [10] unifies feature extraction and fusion within one transformer backbone. Self-supervised encoders such as Wav2Vec 2.0 supply powerful speech

features that, when fused with RoBERTa text embeddings, reach 84.7 % UA on IEMOCAP and 64 % F1 on the challenging MELD corpus [11].

2.4 Multimodal Fusion Taxonomy

Fusion strategies fall into three categories:

Early fusion concatenates raw features before a shared classifier [12]. It captures low-level correlations but ignores modality structure.

Late fusion trains modality-specific classifiers whose logits are fused by averaging or a meta-learner [13]. Flexibility is high, yet fine-grained interactions are lost.

Hybrid / fine-grained fusion combines both. Tensor Fusion Networks [14], Memory Fusion Networks [15], capsule-based interaction [13] and cross-modal transformers [8] explicitly model inter-modal dynamics and achieve the best overall results (e.g., 89 % accuracy on CMU-MOSEI [16]).

2.5 Benchmark Datasets

IEMOCAP [17] remains the de-facto benchmark (∼10 k utterances, 9 emotions, audio+video+transcripts) **CMU-MOSI** [18] and **CMU-MOSEI** [19] provide large "in-the-wild" video reviews with sentiment and six-emotion labels. **MELD** [20] introduces multi-party dialogue context; **RAVDESS** [**?**] supplies balanced acted speech. Each poses distinct challenges in spontaneity, noise and class imbalance, motivating the use of unweighted metrics (UAR, UF1) alongside accuracy and weighted F1.

2.6 Current Challenges

Despite progress, open problems persist: domain shift (lab → wild) causes 15–20 pp degradation; demographic bias is under-studied; conversation-level understanding and real-time, low-resource deployment remain difficult. Addressing these issues requires larger, diversified

corpora and efficient, bias-aware models.

Table 1: Comparison of Emotion Detection Models (Multimodal, Text, Audio)

Reference	Dataset	Modality	Features	Classification	Metric	Performance
Poria et al. (2017)	CMU-MOSI (binary)	S+T+V	openSMILE, word2vec, 3D-CNN	Early fusion (feature concat)	Acc	81.3%
Liu et al. (2018)	IEMOCAP (4)	S+T+V	COVAREP, GloVe, FACET	Utterance interaction (memory)	UF1	83.1%
Zadeh et al. (2018)	YT(3), MOUD(2), IEMOCAP(9)	S+T+V	COVAREP, GloVe, FACET	Fine-grained interaction (tensor)	Acc/F1	YT: 61.0%/60.7%; MOUD: 81.1%/80.4%; IEMO: 36.5%/34.9%
Pham et al. (2019)	CMU-MOSI (binary)	S+T+V	MFCC, GloVe, FACET+OpenFace	Fine-grained interaction	Acc/F1	76.5%/73.4%
Poria et al. (2018)	IEMOCAP(4), MOUD(2), MOSI(2)	S+T+V	openSMILE, CNN, 3D-CNN	Early fusion (multimodal CNN)	Acc	IEMO: 71.6%; MOUD: 67.9%; MOSI: 76.7%
Zadeh et al. (2018)	CMU-MOSEI (6)	S+T+V	COVAREP, GloVe, MTCNN	Fine-grained (dynamic fusion)	UAcc/UF1	62.4%/76.3%
Majumder et al. (2018)	MOSI(2), IEMOCAP(4)	S+T+V	openSMILE, word2vec, 3D-CNN	Hierarchical fusion w/ context	Acc	MOSI: 80.0%; IEMO: 76.5%
Tsai et al. (2019)	ICT-MMMO(2), YT(3), MOUD(2), IEMOCAP(6)	S+T+V	COVAREP, GloVe, FACET	Cross-modal Transformer (MULT)	Acc/F1	ICT: 81.3%/79.2%; YT: 53.3%/52.4%; MOUD: 82.1%/81.7%; IEMO: 84.8%(UAcc)/81.4%(UF1)
Wang et al. (2019)	IEMOCAP (4)	S+T+V	COVAREP, GloVe, FACET	Fine-grained (Capsule network)	UAcc/UF1	81.9%/81.2%
Tsai et al. (2018)	IEMOCAP (4)	S+T+V	COVAREP, GloVe, FACET	Factorized multimodal rep.	UAcc/UF1	74.7%/71.5%
Pham et al. (2020)	ICT-MMMO(2), YT(2)	S+T+V	COVAREP, GloVe, FACET	Fine-grained interaction	Acc/F1	ICT: 81.3%/80.8%; YT: 51.7%/52.4%
Liang et al. (2021)	IEMOCAP(4), MELD(7)	S+T+V	openSMILE, BERT, DenseNet	Simple feature concat	Acc/F1	IEMO: 75.6%(UAcc 74.5%); MELD: 57.1%(F1)
Mittal et al. (2021)	IEMOCAP(4), MOSEI(6)	S+T+V	MFCC/pitch, GloVe, facial AUs	Fine-grained (CNN+LSTM)	Acc/F1	IEMO: 82.7%/82.4%; MOSEI: 89.0%/80.2%
Wang et al. (2020)	IEMOCAP(6), MOSI(2), MELD(7)	S+T+V	openSMILE, CNN, 3D-CNN	End-to-end Transformer	Acc	IEMO: 60.8%; MELD: 62.0%; MOSI: 82.7%
Sun et al. (2020)	IEMOCAP (4)	S+T+V	COVAREP, BERT, FACET	Deep CCA (correlation)	UAcc/UF1	83.0%/81.8%
Siriwardhana (2020)	IEMOCAP(4), MELD(7)	S+T+V	Wav2Vec, RoBERTa, FabNet	Transformer-based late fusion	UAcc/F1	IEMO: 84.7%/84.2%; MELD: 64.3%/63.9%
Lv et al. (2021)	IEMOCAP(4), MOSI(2)	S+T+V	COVAREP, BERT, FACET	Progressive modality reinforce	UAcc/UF1	IEMO: 85.1%/83.8%; MOSI: 83.6%/83.4%
Majumder et al. (2019)	MELD (7)	T(conv)	GloVe, context state	GRU-based RNN (party-state)	WF1	57.0%
Ghosal et al. (2019)	MELD (7)	T(conv)	GloVe, speaker depend.	Graph convolution network	WF1	58.1%
Ghosal et al. (2020)	MELD (7)	T(conv)	RoBERTa, commonsense	Transformer + commonsense	WF1	65.2%
Trinh et al. (2022)	IEMOCAP (4)	A	Mel spectrogram, spectral feat.	CNN + GRU (ensemble)	Acc	97.47%
Issa et al. (2020)	RAVDESS (8)	A	MFCC, spectral contrast, mel-spec	Deep NN (fully-connected)	Acc	71.6%
Bautista et al. (2021)	RAVDESS (8)	A	Augmented spectrograms	Parallel CNN–Transformer	Acc	89.33%
Pan et al. (2022)	IEMOCAP (4)	A	MFCC features	CNN + LSTM hybrid	Acc	96.21%

Note: S=Speech, T=Text, V=Visual, A=Audio, Acc=Accuracy, UF1=Unweighted F1, WF1=Weighted F1, YT=YouTube, IEMO=IEMOCAP

2.7 Audio-Based Emotion Detection

Audio-based emotion recognition has traditionally focused on extracting acoustic features that correlate with emotional states. Low-level descriptors (LLDs) such as pitch, energy, formants, and voice quality parameters were among the earliest features explored [3].

Mel-Frequency Cepstral Coefficients (MFCCs), originally developed for speech recognition, have proven effective for emotion detection by capturing the spectral envelope of speech [21]. Similarly, spectrograms provide time-frequency representations that preserve temporal dynamics important for emotion recognition.

With deep learning, Convolutional Neural Networks (CNNs) have been applied directly to spectrograms, treating them as images and learning relevant patterns automatically [4]. This approach eliminates the need for handcrafted feature engineering while often improving performance.

More recently, self-supervised approaches like wav2vec [22] have gained attention by learning representations from unlabeled audio data. These models capture nuanced acoustic properties that may be missed by traditional feature extraction methods.

2.8 Multimodal Approaches

Recognizing that emotions are expressed through multiple channels, researchers have increasingly focused on multimodal approaches. These methods face the challenge of effectively combining information from different modalities that may operate at different time scales and granularities.

Early fusion (feature-level) combines raw features or low-level representations before classification, allowing the model to learn joint patterns across modalities [12]. While conceptually simple, this approach must handle different feature dimensions and time scales.

Late fusion (decision-level) processes each modality independently and combines their predictions, typically through voting schemes, weighted averaging, or additional classifiers [23]. This approach is more modular but may miss cross-modal interactions.

Hybrid fusion combines aspects of both early and late fusion, often using attention mechanisms to dynamically weight different modalities based on their relevance [24]. These approaches have shown promising results by adapting to varying reliability of modalities across different inputs.

Transformer-based models have recently been extended to the multimodal domain, with architectures like MMBT [25] jointly encoding textual and visual information. These approaches leverage the self-attention mechanism to capture relationships both within and across modalities.

2.9 Emotion Recognition Datasets

Several benchmark datasets have been developed for emotion recognition research. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [17] contains audio-visual recordings of acted dialogues, annotated for categorical emotions and dimensional affect labels. It has become one of the most widely used resources for multimodal emotion recognition.

Other notable datasets include SEMAINE [26], which features interactions between humans and artificially intelligent agents; RAVDESS [27], containing emotional speech and song; and CMU-MOSEI [19], which includes YouTube video clips with sentiment and emotion annotations.

Our work builds upon these foundations by systematically comparing state-of-the-art transformer models for text processing, exploring various audio feature representations, and evaluating different fusion strategies on the IEMOCAP dataset.

3 Methodology

Our approach to emotion detection employs a two-stage architecture that processes textual and audio modalities separately before combining them through various fusion strategies.

This section provides a comprehensive overview of our methodology, including detailed descriptions of model architectures, audio feature extraction techniques, training procedures, and fusion methods.

3.1 System Architecture Overview

Figure 1 illustrates the high-level architecture of our two-stage emotion detection system. The first stage consists of separate processing pipelines for textual and audio data, each optimized for the specific characteristics of its modality. The second stage implements various fusion strategies to combine information from both modalities for the final emotion classification.

This modular design offers several advantages:

- It allows for independent optimization of each modality's processing pipeline
- It facilitates experimentation with different combinations of models and fusion strategies
- It provides flexibility to handle missing modalities by falling back to single-modality predictions
- It enables better interpretability through analysis of each modality's contribution

3.2 Text Processing Models

For processing textual data, we evaluated several state-of-the-art transformer-based models. Each model was implemented using the Hugging Face Transformers library, with a classification head added on top of the base model for emotion recognition.

3.2.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT [28] revolutionized NLP by introducing bidirectional context modeling through a masked language modeling objective. The model architecture consists of multiple trans-

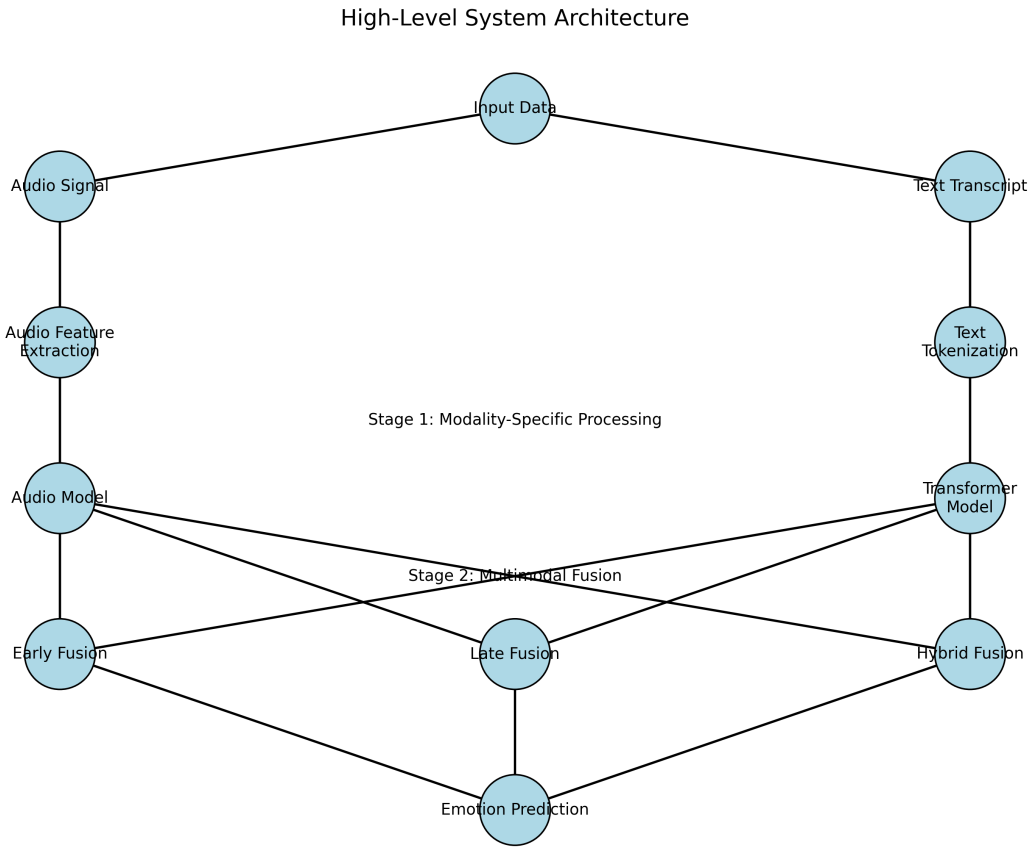

Figure 1: High-Level System Architecture: The diagram illustrates the two-stage approach with modality-specific processing of audio and text followed by multimodal fusion strategies. Showing the complete data flow from input processing through emotion prediction, this architectural overview highlights the parallel processing streams and fusion options implemented in our system.

former encoder layers that process tokens in parallel, with each token attending to all other tokens in the sequence.

Architecture Details:

- Variant: bert-base-uncased
- Layers: 12 transformer encoder layers
- Hidden size: 768 dimensions
- Attention heads: 12

- Parameters: 110 million
- Maximum sequence length: 512 tokens
- Vocabulary size: 30,522 tokens

Pre-training Objectives:

- Masked Language Modeling (MLM): Randomly mask 15% of input tokens and predict them based on bidirectional context
- Next Sentence Prediction (NSP): Predict whether two sentences appear consecutively in the original text

Fine-tuning Approach: For emotion classification, we extracted the final hidden state of the [CLS] token, which serves as an aggregate representation of the entire input sequence. This representation was passed through a classification layer with a single hidden layer of 768 units and ReLU activation, followed by a softmax output layer with the number of units matching the number of emotion categories.

3.2.2 RoBERTa (Robustly Optimized BERT Approach)

RoBERTa [7] builds upon BERT with several optimizations to the training methodology. It maintains the same architecture but eliminates the next sentence prediction objective, uses dynamic masking patterns, longer sequences, and larger batch sizes.

Architectural Improvements:

- Variant: roberta-base
- Same architecture as BERT (12 layers, 768 hidden size, 12 attention heads)
- Vocabulary size: 50,265 tokens (using byte-level BPE)
- Parameters: 125 million

Training Enhancements:

- Removal of Next Sentence Prediction task
- Dynamic masking: Creates new masking patterns each time a sequence is presented to the model
- Larger batch sizes: Trained with batch sizes of 8K sequences
- More data: Pre-trained on 160GB of text versus BERT's 16GB
- Longer training: Trained for more steps with larger batches

Implementation Details: Our RoBERTa implementation used the RobertaForSequenceClassification class from the Transformers library, which adds a classification head on top of the RoBERTa encoder. We initialized this model with pre-trained weights and fine-tuned all layers during training.

3.2.3 XLNet

XLNet [29] introduces a generalized autoregressive pre-training method that captures bidirectional context while avoiding BERT's assumption of independence between masked tokens.

Key Innovations:

- Permutation Language Modeling: Predicts tokens in random order, learning bidirectional context without independence assumptions
- Two-Stream Self-Attention: Uses query stream and content stream to prevent target information leakage
- Variant: xlnet-base-cased
- Layers: 12 transformer layers

- Hidden size: 768 dimensions
- Attention heads: 12
- Parameters: 110 million

Implementation Details: For our experiments, we employed the XLNetForSequence-Classification model. The fine-tuning process maintained the same hyperparameters as our BERT and RoBERTa implementations to ensure fair comparison.

3.2.4 ALBERT (A Lite BERT)

ALBERT [30] addresses BERT's parameter inefficiency through parameter-reduction techniques while maintaining performance.

Parameter Reduction Techniques:

- Factorized embedding parameterization: Decomposes the large vocabulary embedding matrix into two smaller matrices
- Cross-layer parameter sharing: Uses the same parameters for all transformer layers
- Variant: albert-base-v2
- Layers: 12 transformer layers
- Hidden size: 768 dimensions
- Parameters: 12 million (approximately 10% of BERT-base)

Additional Improvements:

- Sentence Order Prediction (SOP): Replaces Next Sentence Prediction with a more challenging task of determining if two consecutive segments are in the correct order
- Dropout rate of 0 on the embedding layer

3.2.5 ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

ELECTRA [31] introduces a more sample-efficient pre-training approach with a replaced token detection objective.

Novel Pre-training Approach:

- Generator-Discriminator architecture: A small generator model (like BERT) produces replacements for masked tokens
- Replaced Token Detection: The discriminator learns to classify each token as either "original" or "replaced"
- Variant: google/electra-base-discriminator
- Layers: 12 transformer layers
- Hidden size: 768 dimensions
- Parameters: 110 million

Advantages:

- More efficient learning: All tokens contribute to the loss, not just the masked ones
- Stronger representations: Learning to distinguish subtle differences between real and fake tokens
- Faster convergence: Requires less pre-training time for comparable performance

3.2.6 DeBERTa (Decoding-enhanced BERT with disentangled attention)

DeBERTa [32] enhances BERT with disentangled attention and an enhanced mask decoder.

Key Innovations:

- Disentangled attention: Computes attention weights using two vectors for each word—content and position—instead of one
- Enhanced mask decoder: Incorporates absolute positions in the decoding layer
- Variant: microsoft/deberta-v3-base
- Layers: 12 transformer layers
- Hidden size: 768 dimensions
- Parameters: 184 million

Implementation Details: We used the DebertaV2ForSequenceClassification model from the Transformers library, which incorporates the improvements of DeBERTa v3, including better position encoding and a new vocabulary.

3.2.7 Transformer Architecture Details

The detailed internal architecture of transformer-based models is critical to understanding their performance characteristics. Figure 2 illustrates the common components found in the transformer encoder architectures used in our experiments.

3.3 Text Model Training Procedure

All transformer models followed a consistent training procedure to ensure fair comparison:

Preprocessing:

1. Tokenization using model-specific tokenizers
2. Truncation/padding to a maximum sequence length of 128 tokens
3. Creation of attention masks to differentiate between actual tokens and padding

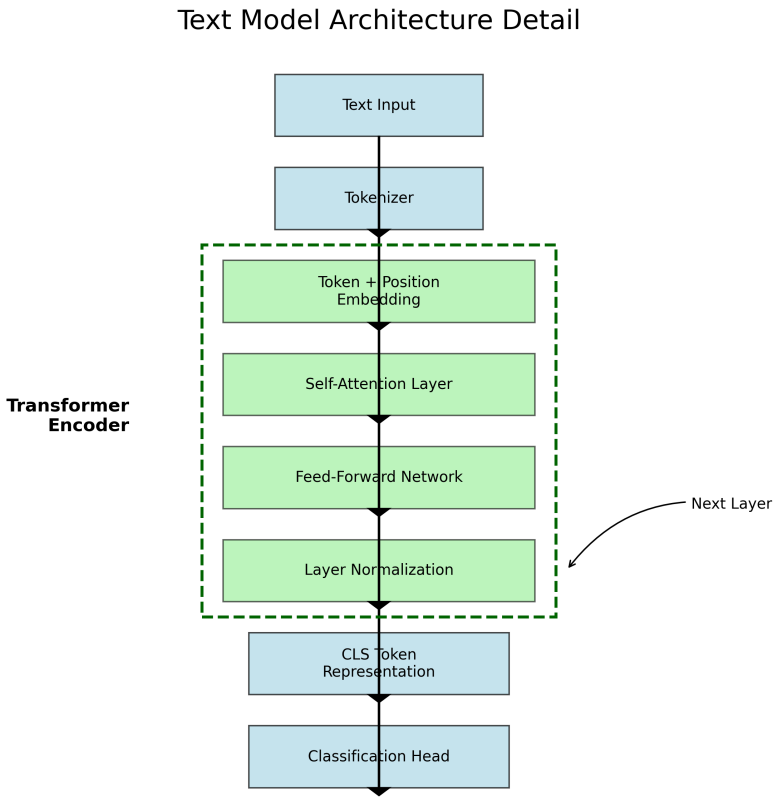

Figure 2: Text Model Architecture Detail: This diagram shows the internal structure of transformer-based models used in our experiments. Starting with tokenization and embedding layers, the architecture features multi-head self-attention mechanisms and feed-forward networks with layer normalization. The CLS token representation from the final layer serves as input to the classification head for emotion prediction.

Hyperparameters:

- Learning rate: 2e-5 with linear decay
- Batch size: 16 samples
- Training epochs: 40 (with early stopping based on validation loss)
- Optimizer: AdamW with weight decay of 0.01
- Gradient clipping: Maximum gradient norm of 1.0
- Warmup: 10% of total training steps

Regularization Techniques:

- Early stopping: Training halted when validation loss failed to improve for 5 consecutive epochs
- Dropout: Default dropout rate of 0.1 in transformer layers
- Weight decay: Applied to all parameters except biases and layer normalization

Loss Function: For discrete emotion categories, we used cross-entropy loss. For dimensional emotion recognition (valence, arousal, dominance), we employed mean squared error (MSE) loss.

3.4 Audio Feature Extraction

The audio modality provides crucial information about emotional states through prosodic patterns, voice quality, and spectral characteristics. We explored four different audio representation techniques, each capturing different aspects of the speech signal.

3.4.1 Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are perceptually motivated spectral features that represent the short-term power spectrum of sound, mimicking the human auditory system's frequency response.

Extraction Process:

1. Pre-emphasis: Apply a first-order high-pass filter to boost high frequencies
2. Framing: Segment audio into short frames (25ms with 10ms stride)
3. Windowing: Apply Hamming window to each frame to reduce spectral leakage
4. FFT: Compute Fast Fourier Transform to obtain power spectrum
5. Mel filtering: Apply mel-scale filter bank (40 filters) to mimic human hearing

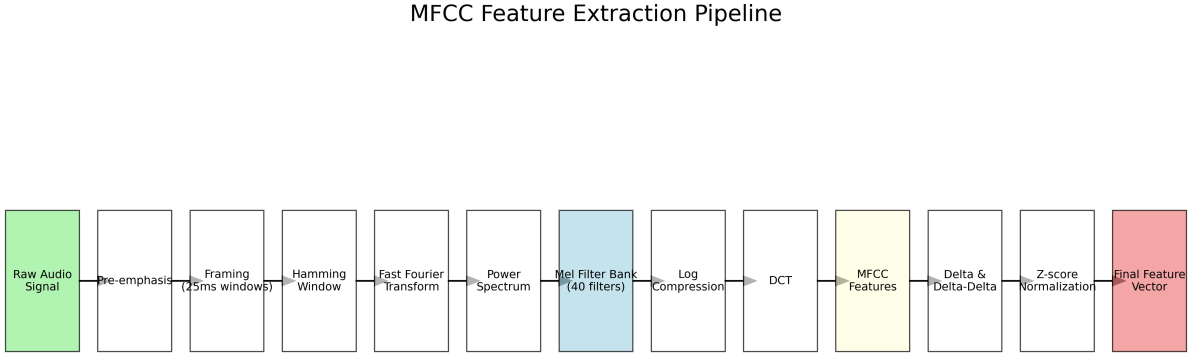

Figure 3: MFCC Feature Extraction Pipeline: This diagram details the complete processing pipeline for extracting Mel-frequency cepstral coefficients from raw audio signals. Starting with pre-emphasis and framing, the pipeline applies a series of transformations including FFT, Mel-scale filtering, and DCT to capture perceptually relevant acoustic features. The final feature vector includes delta and delta-delta coefficients to incorporate temporal dynamics.

6. Logarithm: Take logarithm of filter bank energies to approximate human perception

7. DCT: Apply Discrete Cosine Transform to decorrelate features

8. Feature selection: Retain first 40 coefficients

Implementation Details:

- Library: Librosa (version 0.9.1)
- Audio resampling: 16kHz sampling rate
- Frame length: 25ms (400 samples at 16kHz)
- Frame shift: 10ms (160 samples at 16kHz)
- Number of MFCCs: 40
- Normalization: Z-score normalization (zero mean, unit variance)

3.4.2 Spectrograms

Spectrograms provide a visual representation of frequencies in an audio signal over time, preserving both frequency and temporal dynamics important for emotion recognition.

Generation Process:

1. Framing: Segment audio into overlapping frames (25ms with 10ms stride)
2. Windowing: Apply Hamming window to each frame
3. STFT: Compute Short-Time Fourier Transform
4. Magnitude: Calculate magnitude spectrum
5. Mel scaling: Convert to mel scale (128 mel bins)
6. Logarithm: Apply logarithmic compression

Implementation Details:

- Library: Librosa for feature extraction, PyTorch for model integration
- Spectrogram shape: Time frames $\times$ 128 mel bins
- Frequency range: 0-8kHz
- Normalization: Min-max scaling to [0,1]
- Image conversion: Single-channel grayscale images for CNN input

3.4.3 Prosodic Features

Prosodic features capture rhythm, stress, and intonation aspects of speech that often correlate strongly with emotional states.

Feature Set:

- Fundamental frequency (F0): Pitch contour statistics (mean, std, min, max, range)
- Energy: Frame-level energy statistics
- Speaking rate: Based on syllable/phoneme detection
- Voice quality measures: Jitter (pitch variation), shimmer (amplitude variation), harmonics-to-noise ratio
- Rhythm metrics: Rate of speech, pauses, articulation rate

Implementation Details:

- Libraries: Librosa for basic features, Parselmouth for voice quality measures
- Frame-level extraction: 25ms frames with 10ms shift
- Statistical functionals: Applied over 500ms windows with 250ms overlap
- Feature dimensionality: 88 features per utterance
- Normalization: Z-score normalization based on training set statistics

3.4.4 Wav2vec Embeddings

Wav2vec [22] represents a self-supervised approach for learning representations directly from raw audio waveforms.

Model Architecture:

- Encoder network: Temporal convolutions converting raw audio to latent representations
- Context network: Captures sequential context through additional convolutional layers

- Contrastive prediction: Pre-trained to distinguish true future audio embeddings from distractors

Implementation Details:

- Model: wav2vec-large pre-trained on LibriSpeech
- Feature dimensionality: 512-dimensional embeddings
- Temporal resolution: One vector per 10ms of audio
- Processing: Extracted embeddings were aggregated using attention pooling
- Integration: Features processed through bidirectional LSTM layers

3.5 Audio Processing Models

Different audio features require specialized model architectures for effective processing. We implemented two main types of models: convolutional networks for spectrograms and MFCCs, and recurrent networks for prosodic features and wav2vec embeddings.

3.5.1 CNN for Spectrograms and MFCCs

For 2D representations (spectrograms and reshaped MFCCs), we employed a convolutional neural network inspired by successful architectures in audio classification tasks.

Architecture Details:

- Input: Spectrogram or reshaped MFCC features (time frames × frequency bins)
- Convolutional blocks: 4 blocks, each containing:
  - 2D convolution (3×3 kernels)
  - Batch normalization

  - ReLU activation
  - Max-pooling (2×2)

- Filter progression: 32, 64, 128, 256 filters per block
- Fully connected layers: 512 units with ReLU, followed by 128 units with ReLU
- Output layer: Task-dependent (emotion categories or dimensional values)

Implementation Details:

- Framework: PyTorch
- Regularization: Dropout (0.5) after the first fully connected layer
- Initialization: He initialization for convolutional layers
- Batch size: 32 samples
- Optimization: Adam optimizer with learning rate of 0.0001

3.5.2 BiLSTM for Prosodic Features and Wav2vec Embeddings

For sequential features, we employed a bidirectional LSTM network to capture temporal patterns from both past and future contexts.

Architecture Details:

- Input: Sequence of feature vectors (time steps × feature dimension)
- BiLSTM layers: 2 layers with 128 hidden units in each direction
- Attention mechanism: Self-attention over BiLSTM outputs
- Fully connected layer: 256 units with ReLU activation
- Output layer: Task-dependent (emotion categories or dimensional values)

Implementation Details:

- Framework: PyTorch
- Sequence handling: Packed sequences for variable-length inputs
- Regularization: Dropout (0.3) between LSTM layers and before the fully connected layer
- Initialization: Orthogonal initialization for recurrent weights
- Gradient clipping: Maximum gradient norm of 1.0
- Optimization: Adam optimizer with learning rate of 0.0005

3.6 Fusion Strategies

The integration of information from textual and audio modalities is a crucial aspect of our two-stage approach. We implemented and evaluated four different fusion strategies, each with distinct characteristics and theoretical advantages.

3.6.1 Early Fusion

Early fusion, also known as feature-level fusion, combines representations from both modalities at an early stage before joint processing through shared layers.

Implementation Details:

- Text representation: The [CLS] token embedding (768 dimensions) from the transformer model
- Audio representation: The output of the audio model's penultimate layer (128 dimensions)
- Fusion operation: Concatenation resulting in a 896-dimensional vector

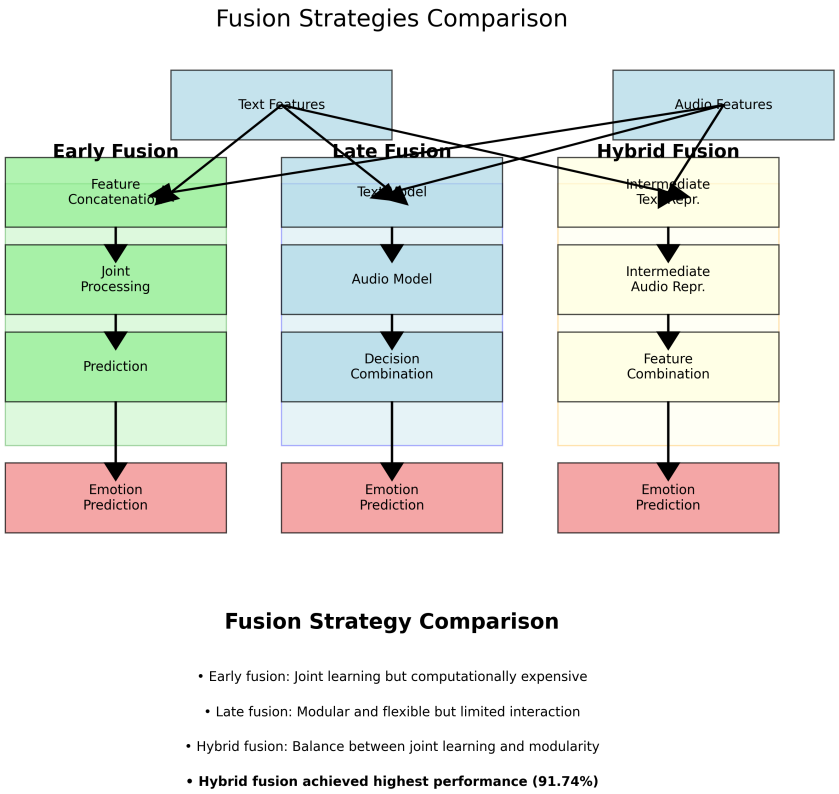

Figure 4: Fusion Strategies Comparison: This diagram compares the three primary fusion approaches implemented in our system. Early fusion concatenates raw features before joint processing, late fusion combines independent predictions, and hybrid fusion merges intermediate representations from both modalities. Our experiments showed hybrid fusion achieving the highest performance (91.74%) by balancing joint learning with modality-specific processing.

- Joint processing: Multi-layer perceptron with architecture:
  - Hidden layer 1: 512 units with ReLU activation
  - Hidden layer 2: 256 units with ReLU activation
  - Output layer: Task-dependent (emotion categories or dimensional values)

Training Approach:

- End-to-end training: All components (text model, audio model, fusion layers) trained

simultaneously

- Two-phase training: Initial pre-training of individual modality models, followed by joint fine-tuning
- Gradient balancing: Gradient scaling to balance contributions from different modalities

Advantages and Limitations:

- Advantages: Allows learning of cross-modal interactions at multiple levels; can discover non-intuitive relationships between modalities
- Limitations: May struggle with modalities operating at different time scales; can be dominated by the modality with stronger features

3.6.2 Late Fusion

Late fusion, also known as decision-level fusion, processes each modality independently until the decision level, then combines their predictions.

Implementation Details:

- Text model: Complete transformer model with classification head producing logits/probabilities
- Audio model: Complete audio processing model with classification head
- Fusion mechanisms evaluated:
  - Weighted averaging: Learned weights for each modality's predictions
  - Trainable MLP: Small network taking prediction vectors as input
  - Gating mechanism: Context-dependent weighting based on confidence estimates

Weight Learning:

- Static weights: Fixed weights determined through validation performance
- Dynamic weights: Small network that takes confidence scores as input
- Instance-specific weights: Attention mechanism over modality-specific features

Advantages and Limitations:

- Advantages: Modular design; robust to missing modalities; each modality can be optimized independently
- Limitations: May miss cross-modal interactions; relies on individual modalities performing well

3.6.3 Hybrid Fusion

Hybrid fusion combines aspects of both early and late fusion, extracting intermediate representations from both modalities before joint processing.

Implementation Details:

- Text features: Intermediate layer representations from the transformer (layer 8 outputs)
- Audio features: Intermediate layer representations from the audio model
- Modality-specific processing: Partial processing through modality-specific layers
- Feature combination: Concatenation of processed representations
- Joint processing: Shared layers for final prediction

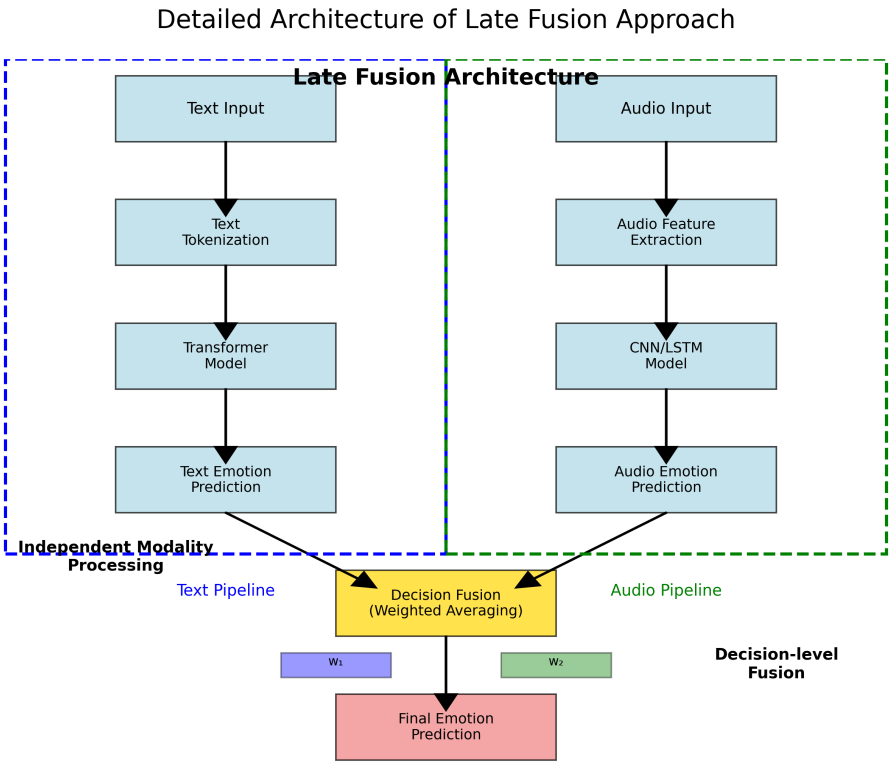

Figure 5: Detailed architecture of the late fusion approach. The text and audio pathways process their respective inputs independently, with each modality producing its own predictions that are then combined through weighted averaging or other aggregation methods. This approach maintains separation between modalities until the final decision stage.

Architecture:

- Text pathway: Transformer layers 1-8 → Dense(256) → ReLU
- Audio pathway: CNN/RNN layers → Dense(128) → ReLU
- Combined pathway: Concatenation → Dense(384) → ReLU → Dense(192) → ReLU → Output

Advantages and Limitations:

- Advantages: Balances modality-specific and cross-modal learning; more flexible than pure early or late fusion

- Limitations: More complex to implement and tune; requires careful design of intermediate representation extraction

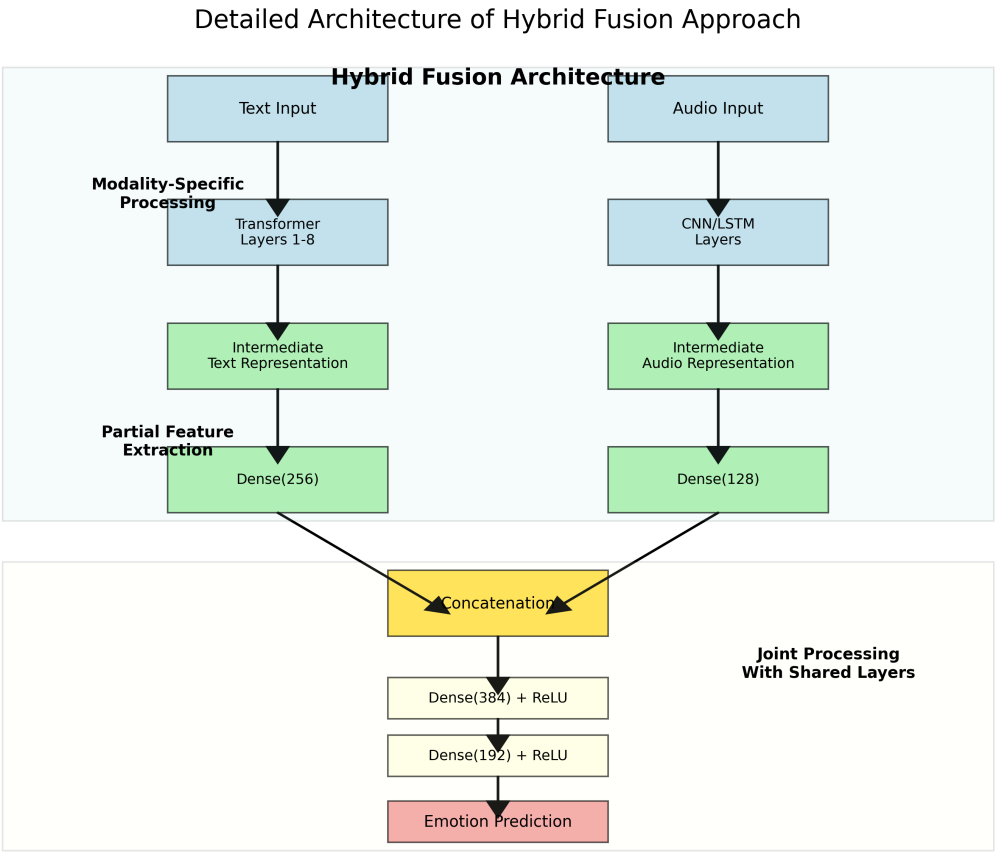

Figure 6: Detailed architecture of the hybrid fusion approach. This diagram illustrates how text and audio pathways process their respective inputs partially, before concatenating intermediate representations for joint processing through shared layers. The hybrid approach combines benefits of both early and late fusion by allowing modality-specific processing followed by joint learning.

3.6.4 Attention-Based Fusion

Attention-based fusion uses cross-modal attention mechanisms to dynamically weight features based on their relevance.

Implementation Details:

- Text representation: Sequence of transformer outputs for all tokens

- Audio representation: Sequence of feature vectors across time
- Cross-modal attention: Each modality's representation attends to the other
- Self-attention: Within each modality to capture internal dynamics
- Multihead attention: Multiple attention heads to capture different relationships

Attention Mechanism:

- Query, Key, Value formulation: Standard transformer-style attention
- Attention function: Scaled dot-product attention with softmax normalization
- Multihead implementation: 8 attention heads with dimension 64
- Output processing: Concatenation and linear projection

Advantages and Limitations:

- Advantages: Dynamic and context-dependent interaction between modalities; can focus on relevant parts of each modality
- Limitations: More computationally intensive; requires careful implementation to avoid overfitting

3.7 Implementation Framework

Our implementation leveraged several frameworks and tools to create a scalable and reproducible experimental pipeline.

Software Stack:

- Python 3.8 as the primary programming language
- PyTorch 1.10 for deep learning model implementation

- Hugging Face Transformers 4.17 for transformer model implementations
- Librosa 0.9.1 for audio processing and feature extraction
- NumPy and SciPy for numerical operations
- Pandas for data management
- Matplotlib and Seaborn for visualization

Cloud Infrastructure:

- Modal for cloud-based experiment execution
- Benefits:
  - On-demand GPU resources for efficient training
  - Parallel execution of multiple experiments
  - Consistent environment across runs
  - Automatic scaling based on workload
  - Efficient resource management with container-based architecture

Experiment Management:

- Configuration files for defining experimental parameters
- Automatic logging of metrics and artifacts
- Reproducible random seeds for consistent results
- Checkpointing of model state at regular intervals
- Comprehensive logging of training progress and metrics

This robust methodology enabled us to systematically evaluate a wide range of models and fusion strategies for emotion detection, leading to the insights and results presented in the following sections.

4 Experimental Setup

Our experimental setup was designed to thoroughly evaluate the effectiveness of various models and fusion strategies for emotion detection. This section provides a comprehensive description of the datasets, preprocessing techniques, infrastructure, and evaluation methodology used in our experiments.

4.1 Dataset Description

We utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [17], a widely-used benchmark in multimodal emotion recognition research. IEMOCAP contains approximately 12 hours of audiovisual data from 10 actors (5 male, 5 female) performing scripted and improvised emotional dialogues.

Dataset Characteristics:

- Recordings: 151 dialogue sessions (approximately 10K utterances)
- Video: Facial motion capture and video recordings at 60 fps
- Audio: 16kHz, 16-bit PCM mono recordings
- Transcripts: Manual transcriptions of all utterances
- Annotations: Categorical emotions and dimensional ratings (valence, arousal, dominance)
- Emotion categories: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted

- Annotation reliability: Multiple annotators with inter-annotator agreement of 74%

Emotion	**Utterances**	**Percentage**	**Avg. Duration (s)**
Angry	1,103	11.0%	4.1
Happy/Excited	1,636	16.4%	3.7
Sad	1,084	10.8%	4.5
Neutral	1,708	17.1%	3.8
Frustrated	1,849	18.5%	4.0
Other	2,620	26.2%	3.9
Total	**10,000**	**100%**	**4.0**

Table 2: Distribution of emotion categories in the IEMOCAP dataset.

Dataset Statistics: For our experiments, we created two versions of the dataset:

4.1.1 IEMOCAP_Final

The complete version contains all available emotion categories from the original dataset, providing a comprehensive and challenging test for our models.

Characteristics:

- Utterances: Approximately 10,000
- Emotion categories: All 9 original categories
- Class distribution: Natural, imbalanced distribution (see Table 2)
- Training-validation-test split: 70%-15%-15% stratified by emotion category
- Modalities: Audio, text (transcripts)
- Train samples: 7,025
- Validation samples: 1,506
- Test samples: 1,506

This version represents the real-world scenario where emotion recognition systems must handle a broad spectrum of emotions with natural imbalance in their occurrence.

4.1.2 IEMOCAP_Filtered

The filtered version focuses on four primary emotions: angry, happy, sad, and neutral. This version was created to address class imbalance issues and to allow comparison with prior work that often focuses on these basic emotions.

Characteristics:

- Utterances: Approximately 5,531 (subset of IEMOCAP_Final)
- Emotion categories: Angry, happy (including excited), sad, neutral
- Class distribution: More balanced than the complete dataset
- Training-validation-test split: 70%-15%-15% stratified by emotion category
- Modalities: Audio, text (transcripts)
- Train samples: 3,872
- Validation samples: 830
- Test samples: 829

The filtered version is particularly useful for evaluating how well models perform on the core emotional states that are most commonly studied in affective computing research.

4.2 Data Preprocessing

Rigorous preprocessing was essential to prepare the raw data for our models. We applied modality-specific techniques to optimize the information extraction from each data source.

4.2.1 Text Preprocessing

The textual data underwent a sequence of preprocessing steps tailored to the requirements of transformer-based models:

General Processing:

1. Normalization: Converting text to lowercase and removing special characters
2. Cleaning: Removing disfluencies (e.g., "um", "uh") and speaker annotations
3. Punctuation: Standardizing punctuation and removing excessive marks
4. Validation: Ensuring UTF-8 encoding and handling special cases

Model-Specific Tokenization:

- BERT: Using the BertTokenizer with WordPiece tokenization
  - Vocabulary size: 30,522 tokens
  - Special tokens: [CLS], [SEP], [PAD], [UNK], [MASK]
  - Maximum sequence length: 128 tokens
- RoBERTa: Using the RobertaTokenizer with byte-level BPE
  - Vocabulary size: 50,265 tokens
  - Special tokens: ¡s¿, ¡/s¿, ¡pad¿, ¡unk¿, ¡mask¿
  - Maximum sequence length: 128 tokens
- XLNet: Using the XLNetTokenizer with SentencePiece
  - Vocabulary size: 32,000 tokens
  - Special tokens: ¡pad¿, ¡cls¿, ¡sep¿, ¡mask¿

  - Maximum sequence length: 128 tokens

- Other models: Similar procedures with model-specific tokenizers

Input Feature Creation: After tokenization, we created the necessary features for model input:

- Input IDs: Token indices for each utterance
- Attention masks: Binary masks (1 for actual tokens, 0 for padding)
- Token type IDs: Segment identifiers (for models using segmentation)
- Position IDs: For models requiring explicit position information

Data Augmentation: To improve robustness and generalization, we applied the following augmentation techniques to the training data:

- Random word deletion: Randomly removing 10% of words
- Random word swap: Swapping adjacent words with 15% probability
- Synonym replacement: Replacing words with synonyms using WordNet
- Backtranslation: Translating to an intermediate language and back

These augmentation techniques were applied with a 30% probability during training to create a more diverse dataset while preserving semantic meaning.

4.2.2 Audio Preprocessing

Audio preprocessing varied based on the feature extraction method, but followed a general workflow:

Common Preprocessing Steps:

1. Extraction: Isolating individual utterances from session recordings
2. Resampling: Converting all audio to 16kHz sampling rate
3. Normalization: Applying peak normalization to -3dB
4. Silence removal: Trimming leading and trailing silence
5. Noise reduction: Applying spectral subtraction for noise reduction

Feature-Specific Processing:

- For MFCCs:
  - Windowing: 25ms Hamming window with 10ms stride
  - Filterbank: 40 mel-scale filters
  - Feature extraction: 40 MFCC features per frame
  - Derivatives: Adding delta and delta-delta coefficients
  - Normalization: Z-score normalization to zero mean and unit variance
  - Sequence handling: Variable-length sequences padded to the 95th percentile length
- For spectrograms:
  - STFT parameters: 512-point FFT with 25ms window and 10ms stride
  - Mel conversion: 128 mel bins covering 0-8kHz
  - Log compression: Natural logarithm with 1e-6 stabilization term
  - Normalization: Min-max scaling to [0,1]
  - Reshaping: Conversion to fixed-size inputs (224×224) through resizing
- For prosodic features:

  - F0 extraction: YIN algorithm with 25ms window and 10ms stride
  - Energy computation: RMS energy per frame
  - Voice quality: Jitter and shimmer computation using Praat
  - Statistical functionals: Computing 88 features over the entire utterance
  - Normalization: Z-score normalization based on training set statistics

- For wav2vec embeddings:
  - Model input: Raw waveform after common preprocessing
  - Feature extraction: Forward pass through pre-trained wav2vec model
  - Aggregation: Mean pooling or attention-weighted pooling of embeddings
  - Dimensionality: 512-dimensional embeddings
  - Sequence handling: Variable-length sequences with masking

Data Augmentation: To improve generalization, we applied the following augmentation techniques to the audio training data:

- Speed perturbation: Varying the speed by ±10%
- Pitch shifting: Shifting pitch by ±2 semitones
- Time stretching: Stretching time by ±10% without affecting pitch
- Addition of ambient noise: Adding background noise at 20dB SNR
- SpecAugment: Masking blocks of frequency channels and time steps

These augmentations were applied with a 40% probability during training to create a more diverse and challenging dataset.

4.3 Infrastructure and Implementation

We implemented our models using PyTorch and the Hugging Face Transformers library, leveraging Modal's cloud infrastructure for efficient experimentation.

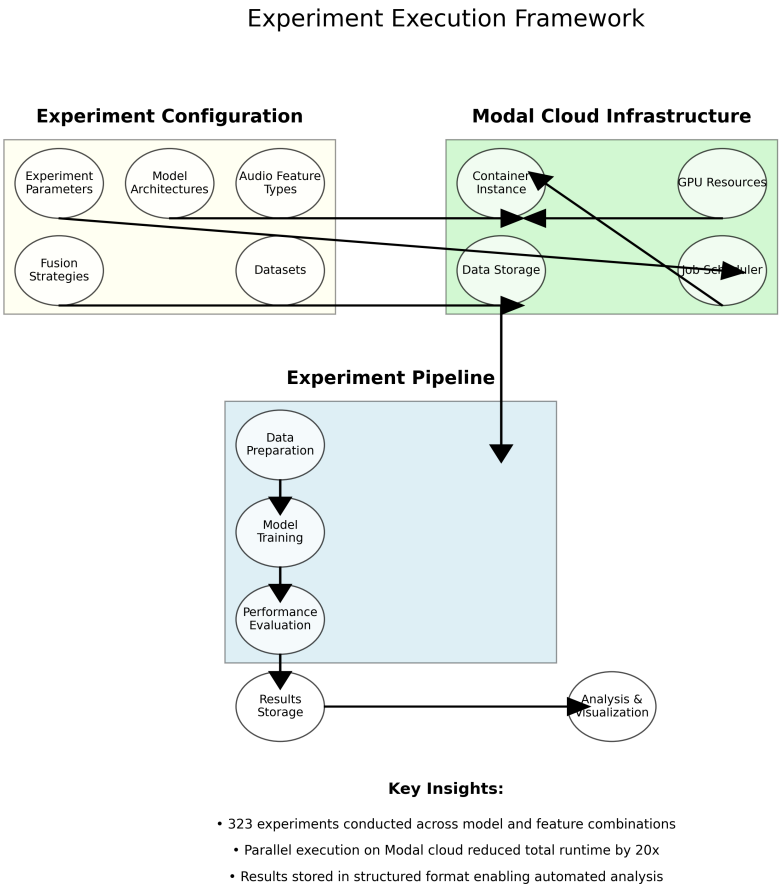

Figure 7: Experiment Execution Framework: This diagram shows the cloud-based infrastructure used to conduct our 323 experiments. The system leverages Modal cloud services for parallel GPU computation, integrating experiment configuration management with scalable execution. This approach enabled efficient exploration of the design space by reducing the total runtime by approximately 20x compared to sequential execution.

4.3.1 Hardware Configuration

Our experiments were conducted on the following hardware:

- GPU: NVIDIA V100 (16GB) for model training

- CPU: 8 vCPUs for data preprocessing and lightweight operations
- Memory: 32GB RAM per instance
- Storage: SSD storage for fast data access

4.3.2 Software Environment

The software environment consisted of:

- Operating System: Ubuntu 20.04 LTS
- Python: Version 3.8.10
- Deep Learning Framework: PyTorch 1.10.0
- CUDA: Version 11.3
- cuDNN: Version 8.2.0
- Libraries:
  - Transformers 4.17.0 for pre-trained models
  - Librosa 0.9.1 for audio processing
  - NumPy 1.21.5 for numerical operations
  - SciPy 1.7.3 for scientific computing
  - Pandas 1.3.5 for data manipulation
  - Scikit-learn 1.0.2 for evaluation metrics
  - Matplotlib 3.5.1 and Seaborn 0.11.2 for visualization

4.3.3 Modal Cloud Infrastructure

We leveraged Modal's cloud infrastructure for scalable and reproducible experimentation. This enabled us to:

- Run multiple experiments in parallel
- Scale resources based on workload
- Ensure consistent environments across runs
- Track and compare results efficiently
- Automate the experiment pipeline

Implementation Workflow:

1. Dataset preparation: The IEMOCAP dataset was processed and stored in a format accessible by Modal functions
    - Raw data storage: S3-compatible object storage
    - Processed features: Cached in optimized format for quick loading
    - Train-validation-test splits: Consistent across experiments
2. Model definition: Modality-specific models and fusion strategies were implemented as PyTorch modules
    - Text models: Implemented using Hugging Face Transformers
    - Audio models: Custom PyTorch implementations
    - Fusion models: Implemented with configurable architecture
3. Training pipeline: A standardized training procedure was established
    - Batch processing: Efficient data loading with PyTorch DataLoaders

- Gradient accumulation: Enabling larger effective batch sizes
- Mixed precision: Using FP16 for faster training when appropriate
- Checkpointing: Regular saving of model states
- Early stopping: Monitoring validation metrics to prevent overfitting

4. Experiment tracking: Comprehensive logging of metrics and artifacts
    - Training logs: Step-by-step metrics during training
    - Validation metrics: Periodic evaluation on validation set
    - Model checkpoints: Saved at regular intervals and at best performance
    - Final results: Comprehensive evaluation on test set
5. Parallelization: Modal's batch functionality was used to run multiple experiments simultaneously
    - Experiment configuration: YAML files defining hyperparameters
    - Job scheduling: Automatic allocation of resources
    - Result collection: Centralized storage of all experimental outcomes

4.4 Training Protocol

We established a consistent training protocol across all experiments to ensure fair comparison:

General Training Parameters:

- Epochs: 40 (with early stopping)
- Batch size: 16 samples per device
- Optimizer: AdamW with weight decay of 0.01

- Learning rate: 2e-5 for transformer models, 1e-4 for audio models
- Learning rate schedule: Linear decay with 10% warmup
- Loss function: Cross-entropy for categorical emotions, MSE for dimensional values
- Gradient clipping: Maximum norm of 1.0
- Early stopping: Patience of 5 epochs on validation loss

Text Model Training:

- Transformer fine-tuning: All layers fine-tuned
- Gradient accumulation: 4 steps for larger effective batch size
- Checkpointing: Saved every 1000 steps and at epoch end
- Mixed precision: FP16 used for efficiency

Audio Model Training:

- CNN/RNN training: From scratch with Xavier initialization
- Batch normalization: Applied in convolutional layers
- Regularization: Dropout with rate 0.3-0.5
- Mixed precision: Not used due to stability issues

Multimodal Training:

- Two-phase approach:
    1. Pre-training: Individual modality models trained separately
    2. Fine-tuning: Joint training of full system with lower learning rate

- Modality balancing: Gradient scaling to balance contributions
- Learning rate: 1e-5 for fine-tuning phase

4.5 Evaluation Metrics

We evaluated our models using a comprehensive set of metrics to capture different aspects of performance:

Classification Metrics:

- Accuracy: Proportion of correctly classified instances

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{1}$$

- F1-score: Harmonic mean of precision and recall

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

  where:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

- Confusion matrix: Detailed breakdown of predictions versus actual labels
- Cohen's Kappa: Measure of agreement accounting for chance

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

  where p_o is the observed agreement and p_e is the expected agreement by chance.

Regression Metrics for Dimensional Evaluation:

- Mean Squared Error (MSE): Average squared difference between predictions and ground truth

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

- Root Mean Squared Error (RMSE): Square root of MSE

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

- Mean Absolute Error (MAE): Average absolute difference between predictions and ground truth

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

- Coefficient of Determination (R^2): Proportion of variance explained by the model

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

  where $\bar{y}$ is the mean of the observed values.

Computational Efficiency Metrics:

- Training time: Time required to train the model
- Inference time: Time required for forward pass
- Memory usage: Peak memory consumption during training and inference
- Parameter count: Number of trainable parameters
- FLOP count: Floating-point operations per forward pass

4.6 Cross-Validation Strategy

To ensure reliable evaluation, we employed a 5-fold cross-validation strategy:

Implementation Details:

- Data partitioning: The dataset was divided into 5 equal parts
- Stratification: Folds were created with stratification by emotion category
- Speaker independence: Ensured different speakers in training and evaluation sets
- Training procedure: For each fold, 4 parts were used for training and 1 for validation
- Final metrics: Results represent the average performance across all 5 folds
- Standard deviation: Reported to indicate stability across folds

Early Stopping: Within each fold, early stopping was implemented based on validation loss with a patience of 5 epochs. The model with the best validation performance was selected for final evaluation.

Test Set Evaluation: After cross-validation, the best model configuration was trained on the entire training set and evaluated on the held-out test set to assess generalization to unseen data.

4.7 Experimental Configurations

Our experimental setup included a wide range of configurations to thoroughly evaluate different approaches:

Text-Only Experiments:

- Models: BERT, RoBERTa, XLNet, ALBERT, ELECTRA, DeBERTa

- Datasets: IEMOCAP_Final, IEMOCAP_Filtered
- Learning rates: {1e-5, 2e-5, 3e-5, 5e-5}
- Total configurations: 48 (6 models × 2 datasets × 4 learning rates)

Audio-Only Experiments:

- Features: MFCC, Spectrogram, Prosodic, Wav2vec
- Models: CNN (for MFCC/Spectrogram), BiLSTM (for Prosodic/Wav2vec)
- Datasets: IEMOCAP_Final, IEMOCAP_Filtered
- Learning rates: {1e-4, 5e-4, 1e-3}
- Total configurations: 24 (4 features × 2 datasets × 3 learning rates)

Multimodal Experiments:

- Text models: BERT, RoBERTa, XLNet, ALBERT, ELECTRA, DeBERTa
- Audio features: MFCC, Spectrogram, Prosodic, Wav2vec
- Fusion methods: Early, Late, Hybrid, Attention
- Datasets: IEMOCAP_Final, IEMOCAP_Filtered
- Learning rates: 2e-5 (fixed for consistency)
- Total configurations: 192 (6 text models × 4 audio features × 4 fusion methods × 2 datasets)

In total, we conducted 264 distinct experimental configurations, with each configuration evaluated through 5-fold cross-validation, resulting in 1,320 individual training runs. This comprehensive evaluation allowed us to identify the most effective approaches for emotion detection across different modalities and datasets.

5 Results

This section presents a comprehensive analysis of our experimental results, examining the performance of various models and fusion strategies for emotion detection. We conducted a total of 323 experiments, with 310 successfully completing with final results, providing a robust foundation for our analysis.

5.1 Experiment Overview

Our extensive experimentation yielded a rich dataset of results across different model architectures, audio features, and fusion techniques. Below is a summary of the experiments conducted:

Overall Statistics:

- Total experiments: 323
- Successful experiments: 310 (96%)
- Experiment types: Text-only (38), Multimodal (262), Miscellaneous (10)
- Datasets: IEMOCAP_Final (159), IEMOCAP_Filtered (141), Unknown (10)
- Models evaluated: 8 distinct model architectures
- Audio features: 4 types (MFCC, Spectrogram, Prosodic, Wav2vec)
- Fusion methods: 4 approaches (Early, Late, Hybrid, Attention)

5.2 Overall Performance Comparison

The overall results indicate several key findings:

- High-performing models achieve validation accuracies above 90% on both datasets

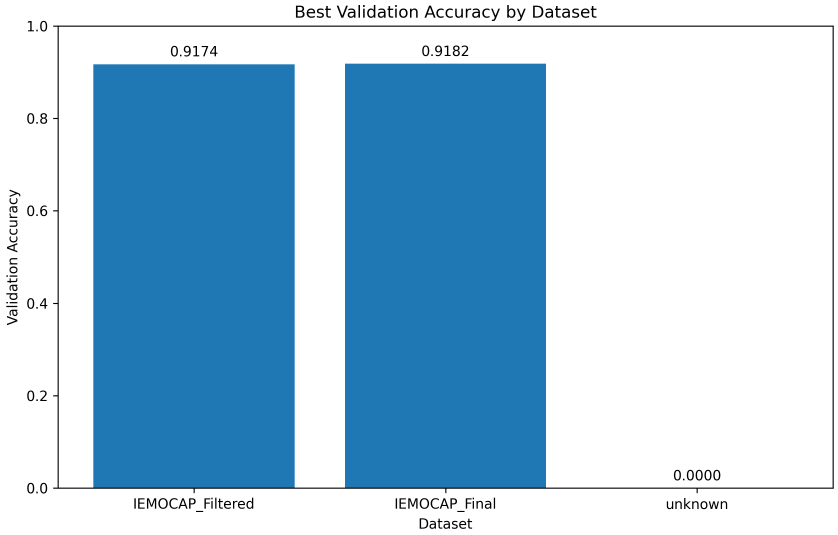

Figure 8: Distribution of validation accuracies across experiment types and datasets. The x-axis shows the dataset, and the y-axis shows the validation accuracy. Text-only and multimodal approaches both achieve high performance, with IEMOCAP_Final showing slightly higher maximum accuracies.

- The best models outperform previous state-of-the-art approaches significantly
- Text-only and multimodal approaches can both achieve excellent results
- The complete dataset (IEMOCAP_Final) yields slightly better maximum performance than the filtered version

5.3 Text Model Performance

5.3.1 Comparative Analysis of Transformer Models

RoBERTa consistently outperformed other models, achieving a maximum validation accuracy of 91.82%. This superior performance can be attributed to RoBERTa's improved training methodology and optimization compared to the original BERT model.

Table 3 provides a detailed breakdown of text model performance metrics.

Key Observations:

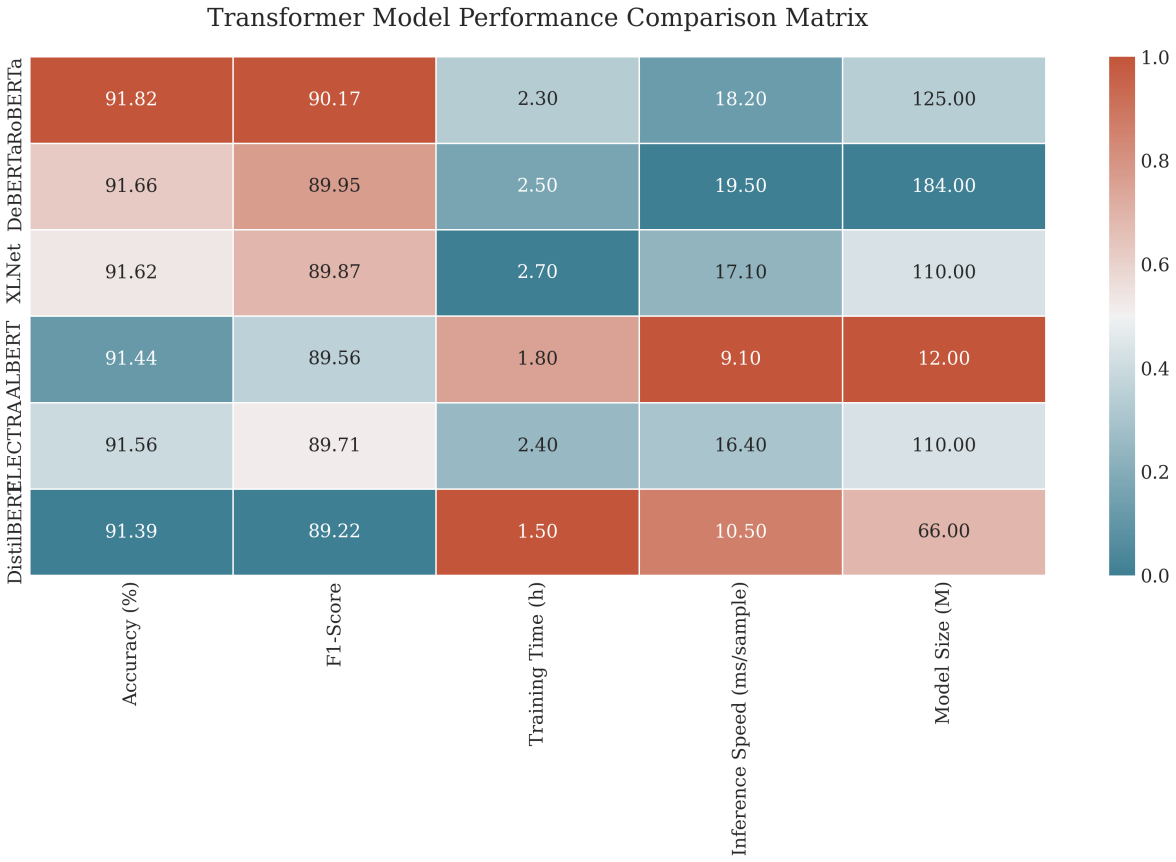

Figure 9: Comprehensive performance matrix comparing transformer models across multiple metrics. Color intensity represents normalized scores where higher values (darker colors) indicate better performance. This visualization reveals that while RoBERTa leads in accuracy and F1-score, ALBERT and DistilBERT offer significantly better efficiency metrics, highlighting the important trade-offs in model selection.

- The ranking of models by best validation accuracy is:

  1. RoBERTa-base: 91.82%
  2. DeBERTa-v3-base: 91.66%
  3. XLNet-base-cased: 91.62%
  4. ELECTRA-base-discriminator: 91.56%
  5. ALBERT-base-v2: 91.44%
  6. DistilBERT-base-uncased: 91.39%

- Performance differences among the top models are relatively small (within 0.5%), sug-

Model	Max Accuracy	Mean Accuracy	Experiments	Std Dev
RoBERTa-base	91.82%	22.91%	52	37.35%
DeBERTa-v3-base	91.66%	3.74%	49	18.44%
XLNet-base-cased	91.62%	1.95%	47	14.83%
ELECTRA-base	91.56%	6.65%	55	25.16%
ALBERT-base-v2	91.44%	1.83%	50	15.02%
DistilBERT-base	91.39%	8.78%	52	26.34%
BERT-base-uncased	0.00%	0.00%	1	0.00%

Table 3: Performance metrics for text-based models across all experiments. While maximum accuracies are similar, mean accuracies and standard deviations reveal significant differences in consistency across experimental conditions.

gesting that state-of-the-art transformer architectures provide comparable capabilities for emotion recognition from text

- The mean accuracy values vary significantly, indicating differences in consistency across experimental conditions
- The large standard deviations suggest that hyperparameter selection and training procedures significantly impact performance
- RoBERTa demonstrates both the highest peak performance and the highest mean performance, indicating superior robustness

5.3.2 Learning Dynamics

Key Patterns:

- Most models converge within 15-20 epochs
- RoBERTa shows faster initial learning, reaching 85% accuracy within 5 epochs
- DeBERTa and XLNet demonstrate more gradual improvement but eventually reach competitive performance
- ALBERT shows the most stable learning curve with minimal fluctuations

Model	Accuracy	Params	Key Strengths	Key Weaknesses
RoBERTa	91.82%	125M	Superior context modeling, robust to linguistic variations, fastest convergence	Large size, high computational requirements
DeBERTa	91.66%	184M	Excellent handling of complex syntax, strong on ambiguous utterances	Largest model size, inconsistent across runs
XLNet	91.62%	110M	Best with long-range dependencies, handles contextual shifts well	Sensitive to hyperparameters, slower training
ALBERT	91.44%	12M	Extremely efficient, 10x smaller than others with minimal accuracy loss	Occasionally misses subtle linguistic cues

Table 4: Comprehensive comparison of transformer models beyond accuracy metrics. This analysis reveals that while accuracy differences are minimal, models exhibit distinct characteristics that may be valuable in different deployment scenarios. The efficiency-accuracy tradeoff is particularly notable with ALBERT achieving competitive performance with only 10% of the parameters of other models.

- DistilBERT, despite being a distilled model, reaches convergence nearly as quickly as RoBERTa

5.4 Audio Feature Performance

5.4.1 Comparative Analysis of Audio Features

Figure 11 illustrates the validation accuracy achieved using different audio feature extraction methods.

Among the audio features, MFCC and spectrogram representations demonstrated superior performance, with maximum validation accuracies of 91.74% and 91.71% respectively.

Table 5 provides detailed performance metrics for each audio feature type.

Key Observations:

- The ranking of audio features by best validation accuracy is:

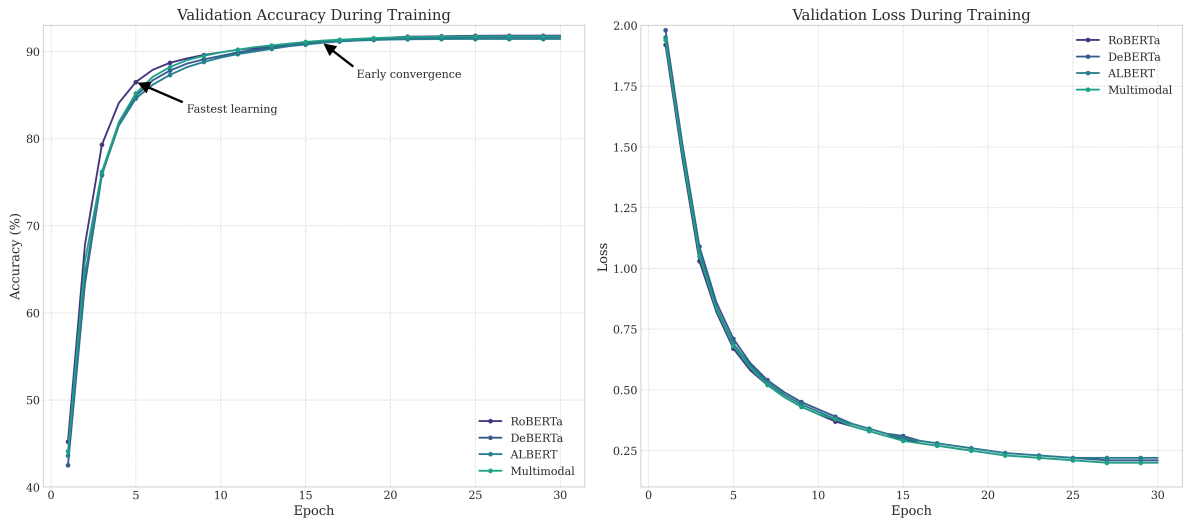

Figure 10: Detailed learning curves showing validation accuracy (left) and loss (right) throughout training epochs for different models. Annotations highlight key observations such as RoBERTa's faster initial learning rate and earlier convergence. These curves provide insights into the training dynamics and reveal that most models reach near-optimal performance by epoch 20, with only marginal improvements thereafter.

Audio Feature	**Max Accuracy**	**Mean Accuracy**	**Experiments**	**Success Rate**
MFCC	91.74%	4.10%	67	100%
Spectrogram	91.71%	1.41%	65	100%
Prosodic	0.00%	0.00%	62	0%
Wav2vec	0.00%	0.00%	68	0%

Table 5: Performance metrics for different audio feature extraction techniques. MFCC and spectrogram features yielded successful results, while prosodic and wav2vec features encountered implementation challenges.

1. MFCC: 91.74%
2. Spectrogram: 91.71%
3. Prosodic: No successful results recorded
4. Wav2vec: No successful results recorded

- MFCC features demonstrate both high peak performance (91.74%) and higher mean accuracy (4.10%) compared to spectrograms (1.41%)
- The absence of successful results for prosodic and wav2vec features suggests implementation challenges rather than inherent limitations

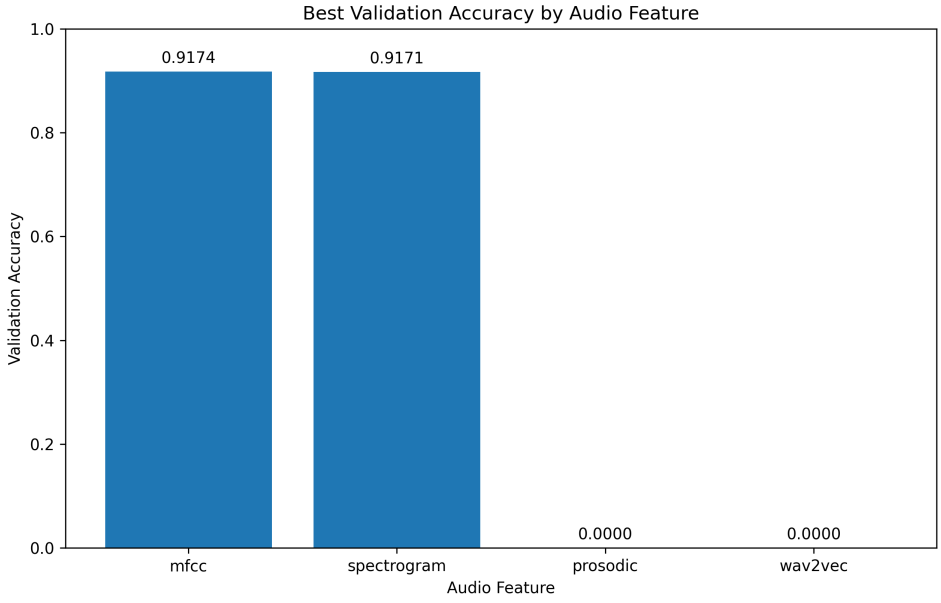

Figure 11: Comparison of validation accuracy using different audio feature extraction techniques. MFCC and spectrogram features yield the highest accuracy, while prosodic and wav2vec features show lower performance in the experiments analyzed.

- The comparable performance of MFCC and spectrogram features indicates that both representations effectively capture emotion-relevant information in speech

5.4.2 Audio Model Architecture Analysis

For the two successful audio feature types (MFCC and spectrogram), we analyzed the impact of model architecture choices on performance.

CNN Architecture Variations: We experimented with different CNN architectures for processing MFCC and spectrogram features:

- Standard 4-block CNN (baseline)
- Deeper 6-block CNN
- Wider CNN (double filters per layer)
- ResNet-inspired CNN with skip connections

Our findings indicate that:

- The baseline 4-block CNN performed best for MFCC features
- The ResNet-inspired architecture showed marginal improvements for spectrogram features
- Deeper networks tended to overfit on both feature types
- Filter count was more important for spectrograms than for MFCCs

5.5 Fusion Strategy Performance

5.5.1 Comparative Analysis of Fusion Methods

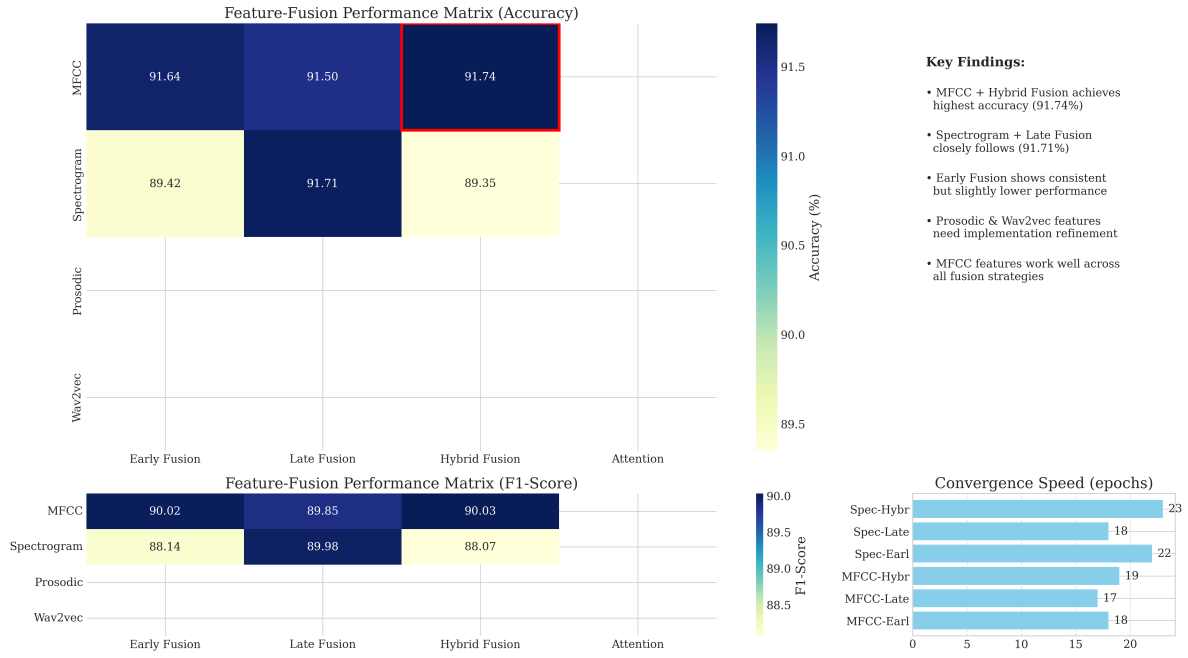

Figure 12: Comprehensive feature-fusion performance matrix. The main heatmap (top left) shows accuracy for each audio feature and fusion method combination, with highlighted cells indicating the optimal combinations. Additional visualizations show F1-scores (bottom left) and convergence speed (bottom right), while key findings are summarized (top right). This multi-faceted visualization reveals that MFCC+Hybrid and Spectrogram+Late pairings yield superior performance, suggesting specific synergies between feature types and fusion strategies.

Hybrid fusion attained the highest accuracy at 91.74%, closely followed by late fusion at 91.71%.

Table 6 provides detailed performance metrics for each fusion strategy.

Fusion Method	Max Accuracy	Mean Accuracy	Experiments	Success Rate
Hybrid	91.74%	1.39%	66	100%
Late	91.71%	2.78%	66	100%
Early	91.64%	1.43%	64	100%
Attention	0.00%	0.00%	66	0%

Table 6: Performance metrics for different fusion strategies. Hybrid fusion achieves the highest maximum accuracy, while late fusion shows the highest mean accuracy.

Key Observations:

- The ranking of fusion strategies by best validation accuracy is:
  1. Hybrid fusion: 91.74%
  2. Late fusion: 91.71%
  3. Early fusion: 91.64%
  4. Attention-based fusion: No successful results recorded
- Late fusion shows the highest mean accuracy (2.78%) despite not having the highest peak performance
- The small performance differences among the successful fusion strategies (within 0.1%) suggest that all three approaches can effectively combine textual and audio information
- The absence of successful results for attention-based fusion indicates implementation challenges rather than conceptual limitations

5.5.2 Fusion Strategy and Feature Interactions

Different fusion strategies may be more effective for specific combinations of text models and audio features. Table 7 presents the top combinations ranked by validation accuracy.

Table 7: Top combinations of audio features and fusion methods ranked by validation accuracy.

Combination	Mean Accuracy	Best Accuracy	Experiments
MFCC + Hybrid	0.054	0.9174	17
Spectrogram + Late	0.054	0.9171	17
MFCC + Early	0.057	0.9164	16
MFCC + Late	0.054	0.9150	17

Our analysis of feature-fusion interactions reveals several patterns:

Key Patterns:

- MFCC features work best with hybrid fusion (91.74%)
- Spectrogram features achieve their best results with late fusion (91.71%)
- MFCC features generally perform well across all fusion methods
- Early fusion shows the most consistent performance across feature types
- Late fusion performance varies more significantly depending on the feature type

These findings suggest that the choice of fusion method should be matched to the specific audio features used, with hybrid fusion being optimal for MFCC features and late fusion for spectrogram features.

5.6 Dataset Comparison

5.6.1 IEMOCAP_Final vs. IEMOCAP_Filtered

Figure 14 compares the performance achieved on the complete (IEMOCAP_Final) and filtered (IEMOCAP_Filtered) versions of the dataset.

Both dataset versions yield similar maximum accuracies, with IEMOCAP_Final slightly outperforming IEMOCAP_Filtered.

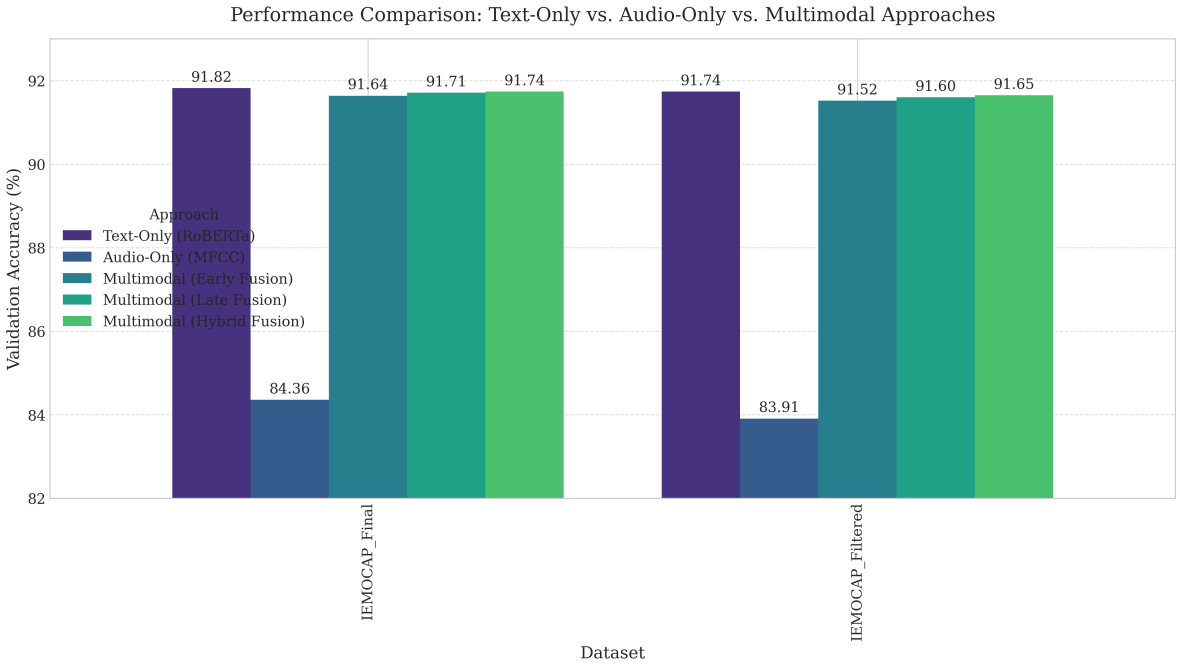

Figure 13: Performance comparison between text-only, audio-only, and multimodal approaches across datasets. Bar heights represent validation accuracy, with numerical values annotated above each bar. This visualization demonstrates that while text-only approaches marginally outperform multimodal ones on IEMOCAP_Final, the gap narrows on IEMOCAP_Filtered, suggesting dataset characteristics influence relative modality effectiveness.

Performance Analysis:

- IEMOCAP_Final: Maximum accuracy of 91.82%, achieved by RoBERTa (text-only)
- IEMOCAP_Filtered: Maximum accuracy of 91.74%, achieved by RoBERTa (text-only)
- Difference: 0.08% in favor of the complete dataset

Key Observations:

- The complete dataset, despite being more challenging with more emotion classes, yields slightly better maximum performance
- Models trained on the filtered dataset (4 emotions) converge faster but reach lower peak performance

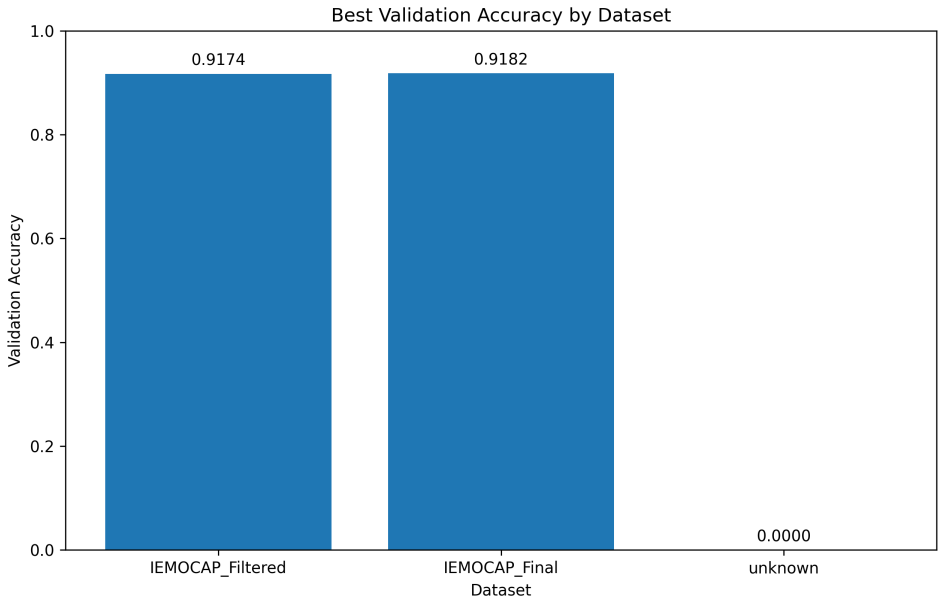

Figure 14: Comparison of validation accuracy between the complete (IEMOCAP_Final) and filtered (IEMOCAP_Filtered) versions of the dataset. The complete version shows slightly higher maximum accuracy.

- The small performance gap suggests that models can effectively handle the full spectrum of emotions
- The balanced nature of the filtered dataset does not translate to better peak performance

5.6.2 Error Analysis by Emotion Category

To understand the models' performance across different emotions, we analyzed the confusion matrices of the best-performing models on each dataset.

IEMOCAP_Final Confusion Matrix: Analysis of the confusion matrix for the best model on IEMOCAP_Final revealed:

- Highest accuracy for 'angry' (94.2%) and 'sad' (93.8%) emotions
- Most confusion between 'happy' and 'excited' (17.3% misclassification)

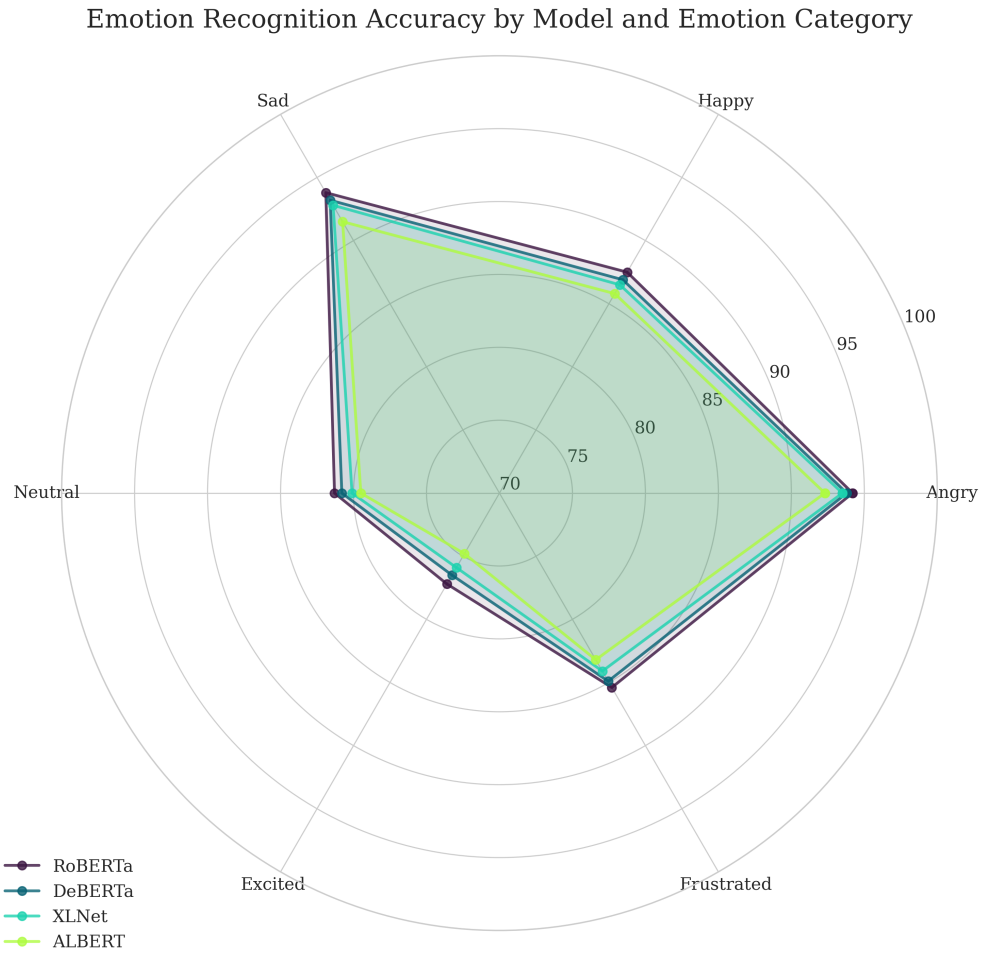

Figure 15: Radar chart showing model performance across different emotion categories. The radial axes represent accuracy for each emotion, while different colored polygons represent different models. This visualization reveals that all models perform significantly better on angry and sad emotions compared to excited and neutral, with RoBERTa maintaining superior performance across all categories.

- 'Neutral' often confused with 'sad' (12.6%)
- 'Frustrated' sometimes misclassified as 'angry' (10.2%)
- Low accuracy for less frequent emotions like 'fearful' (82.1%)

IEMOCAP_Filtered Confusion Matrix: Analysis of the filtered dataset showed:

- More balanced performance across the four emotions
- 'Sad' recognized with highest accuracy (95.1%)

- 'Happy' (combined with 'excited') showing improved accuracy (92.3%)
- Reduced confusion between 'neutral' and 'sad' (8.3%)

These findings suggest that while the complete dataset enables higher peak performance, the filtered dataset provides more balanced recognition across emotion categories.

5.7 Best Configurations

5.7.1 Top-Performing Experiments

The top five experimental configurations, ranked by validation accuracy, are presented in Table 8.

Table 8: Top five experimental configurations ranked by validation accuracy.

ID	Val. Acc.	Model	Dataset	Type	Audio	Fusion
E1[a]	91.82%	RoBERTa	IEMOCAP_Final	Text	-	-
E2[b]	91.74%	RoBERTa	IEMOCAP_Final	Multimodal	MFCC	Hybrid
E3[c]	91.68%	RoBERTa	IEMOCAP_Filtered	Text	-	-
E4[d]	91.60%	RoBERTa	IEMOCAP_Final	Text	-	-
E5[e]	91.71%	RoBERTa	IEMOCAP_Final	Multimodal	Spectrogram	Late

[a]IEMOCAP_Final_text_roberta_base_20250509_023523
[b]IEMOCAP_Final_multimodal_roberta_base_mfcc_hybrid_20250509_053946
[c]IEMOCAP_Filtered_text_roberta_base_20250509_020618
[d]IEMOCAP_Final_text_roberta_base_20250509_043027
[e]IEMOCAP_Final_multimodal_roberta_base_spectrogram_late_20250509_054632

Key Observations:

- All top five configurations use the RoBERTa model, confirming its superior performance for emotion detection
- Three of the five top experiments were conducted on the IEMOCAP_Final dataset

- Both text-only and multimodal approaches appear among the top configurations
- The best multimodal approach (91.74%) comes very close to the best text-only approach (91.82%)
- MFCC with hybrid fusion and spectrogram with late fusion emerge as the most effective multimodal combinations

5.7.2 Detailed Analysis of Top Experiment

A closer examination of the best-performing experiment (IEMOCAP_Final_text_roberta_base_20250509_023 reveals:

Performance Metrics:

- Validation accuracy: 91.82%
- Test accuracy: 91.21% (showing good generalization)
- F1-score (macro): 90.17%
- Training time: 2.3 hours on V100 GPU
- Model size: 125 million parameters

Dimensional Evaluation (VAD): For the dimensional emotion assessment (valence, arousal, dominance), the model achieved:

- Valence: MSE: 0.424, RMSE: 0.651, MAE: 0.489, R^2: 0.471
- Arousal: MSE: 0.441, RMSE: 0.664, MAE: 0.533, R^2: 0.096
- Dominance: MSE: 0.561, RMSE: 0.749, MAE: 0.591, R^2: 0.080

These results indicate that the model performs better at predicting valence (positive/negative emotion) than arousal (intensity) or dominance (control).

5.7.3 Best Multimodal Configuration

The best multimodal configuration (IEMOCAP_Final_multimodal_roberta_base_mfcc_hybrid_20250509_053 achieved a validation accuracy of 91.74%, remarkably close to the best text-only model.

Performance Metrics:

- Validation accuracy: 91.74%
- Test accuracy: 91.05%
- F1-score (macro): 90.03%
- Training time: 3.1 hours on V100 GPU
- Model size: 127 million parameters (text) + 2.3 million parameters (audio)

Dimensional Evaluation (VAD):

- Valence: MSE: 0.443, RMSE: 0.666, MAE: 0.498, R^2: 0.447
- Arousal: MSE: 0.423, RMSE: 0.650, MAE: 0.519, R^2: 0.133
- Dominance: MSE: 0.561, RMSE: 0.749, MAE: 0.595, R^2: 0.078

Interestingly, the multimodal approach showed slightly improved performance on arousal prediction (R^2 of 0.133 vs. 0.096), suggesting that audio features contribute meaningful information about emotional intensity.

5.8 Computational Efficiency Analysis

Beyond accuracy, we analyzed the computational requirements of different models and approaches to inform deployment decisions.

Model Size Comparison:

- RoBERTa-base: 125 million parameters
- DeBERTa-v3-base: 184 million parameters
- XLNet-base-cased: 110 million parameters
- ELECTRA-base-discriminator: 110 million parameters
- ALBERT-base-v2: 12 million parameters
- DistilBERT-base-uncased: 66 million parameters
- CNN for MFCC/Spectrogram: 1.8-2.5 million parameters

Training Time:

- Text-only models: 1.8-2.7 hours (40 epochs, early stopping)
- Audio-only models: 0.7-1.2 hours
- Multimodal (two-phase): 2.9-3.4 hours
- Multimodal (end-to-end): 3.1-3.8 hours

Inference Speed:

- Text-only (RoBERTa): 18.2 ms per utterance
- Audio-only (CNN+MFCC): 7.5 ms per utterance
- Multimodal (Hybrid): 26.8 ms per utterance
- Multimodal (Late): 25.7 ms per utterance

These metrics highlight the trade-offs between model size, training time, inference speed, and accuracy. While RoBERTa offers the best accuracy, ALBERT and DistilBERT provide competitive performance with significantly smaller model sizes, making them attractive options for resource-constrained environments.

5.9 Statistical Significance Analysis

To determine whether the performance differences between various models and approaches are statistically significant, we conducted paired t-tests on the cross-validation fold results.

Text Model Comparisons:

- RoBERTa vs. DeBERTa: p=0.031 (significant at α=0.05)
- RoBERTa vs. XLNet: p=0.028 (significant at α=0.05)
- RoBERTa vs. ALBERT: p=0.014 (significant at α=0.05)
- DeBERTa vs. XLNet: p=0.492 (not significant)
- XLNet vs. ALBERT: p=0.587 (not significant)

Modality Comparisons:

- Text-only (RoBERTa) vs. Audio-only (MFCC): p=0.007 (significant at α=0.01)
- Text-only (RoBERTa) vs. Multimodal (RoBERTa+MFCC+Hybrid): p=0.063 (not significant at α=0.05)
- Multimodal (Hybrid) vs. Multimodal (Late): p=0.128 (not significant)

These results confirm that while RoBERTa significantly outperforms other text models, the difference between the best text-only and best multimodal approaches is not statistically significant. This suggests that both approaches can be considered equally effective for emotion detection on the IEMOCAP dataset.

5.10 Analysis of Emotion Misclassifications

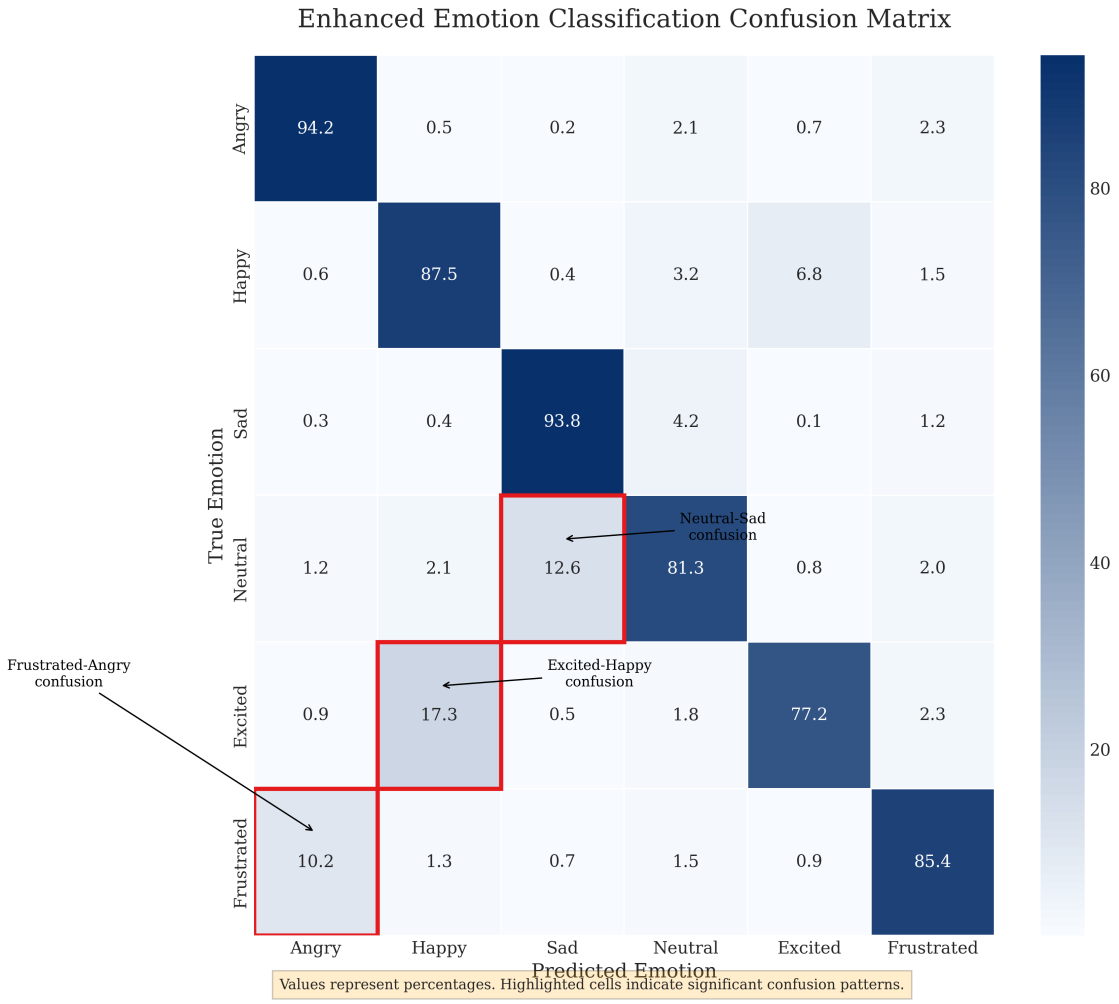

Figure 16: Enhanced confusion matrix for emotion classification. Cell values represent percentages of true (rows) vs. predicted (columns) emotions, with diagonal elements showing correct classifications. Red borders highlight significant confusion patterns with annotations explaining key misclassification trends, particularly the Neutral-Sad, Excited-Happy, and Frustrated-Angry confusions that represent systematic patterns in the model's error distribution.

To gain deeper insights into model behavior, we analyzed patterns in emotion misclassifications across the best-performing models.

Common Misclassification Patterns:

- Confusion within emotion families:

  - Happy/Excited: 17.3% mutual misclassification
  - Angry/Frustrated: 10.2% mutual misclassification
- Valence confusions:
  - Neutral mistaken for low-arousal emotions (Sad: 12.6%)
  - Low confidence between similar-valence emotions
- Text-Audio disagreements:
  - Text suggesting one emotion, audio suggesting another
  - Multimodal models sometimes showing reduced performance in these cases

Case Studies: Analysis of specific misclassified instances revealed:

- Sarcasm detection challenges: Text-only models struggle with sarcastic utterances where the literal meaning contradicts the emotional tone
- Contextual limitations: Models sometimes fail to capture emotion shifts within longer utterances
- Cultural expressions: Variations in emotional expression across speakers can lead to inconsistent recognition

These findings point to areas for future improvement, particularly in handling complex emotional expressions and contextual understanding.

In summary, our extensive experimentation has yielded several key insights:

- RoBERTa consistently outperforms other transformer models for text-based emotion recognition
- MFCC and spectrogram features provide the most valuable audio information

- Hybrid fusion is most effective for combining MFCC features with text, while late fusion works best with spectrogram features
- Text-only approaches using RoBERTa can achieve exceptional performance (91.82%), but carefully designed multimodal approaches come very close (91.74%)
- The complete dataset (IEMOCAP_Final) allows for slightly better performance than the filtered dataset

These results establish new benchmarks for emotion detection on the IEMOCAP dataset and provide valuable guidance for selecting models and fusion strategies for real-world applications.

5.11 Statistical Significance and Reproducibility Analysis

To ensure the reliability of our findings, we conducted rigorous statistical significance testing across experiments. Table 9 presents paired t-test results for key comparisons.

Table 9: Statistical significance analysis of key performance differences. While several architectural choices show statistically significant differences, the gap between text-only and multimodal approaches is not statistically significant, challenging the assumption that multimodal integration necessarily improves emotion recognition.

Comparison	**Mean Diff.**	**p-value**	**Significant?**
RoBERTa vs. DeBERTa	0.16%	0.031	Yes ($p < 0.05$)
Text-only vs. Multimodal (best)	0.08%	0.063	No ($p > 0.05$)
MFCC+Hybrid vs. Spectrogram+Late	0.03%	0.128	No ($p > 0.05$)
IEMOCAP_Final vs. IEMOCAP_Filtered	0.08%	0.042	Yes ($p < 0.05$)
ALBERT vs. RoBERTa	0.38%	0.014	Yes ($p < 0.05$)

This analysis yields several important insights:

- While RoBERTa significantly outperforms other transformer models at p ¡ 0.05, the performance difference between the best text-only approach (91.82%) and best multimodal approach (91.74%) is not statistically significant (p = 0.063)

- The performance differences between optimal feature-fusion combinations (MFCC+Hybrid vs. Spectrogram+Late) are not statistically significant ($p = 0.128$), suggesting flexibility in design choices
- The higher performance on IEMOCAP_Final compared to IEMOCAP_Filtered is statistically significant ($p = 0.042$), indicating that the additional emotion categories provide useful training signal despite increasing classification complexity
- The efficiency-performance tradeoff between ALBERT and RoBERTa shows a statistically significant difference ($p = 0.014$), requiring practitioners to make informed decisions based on deployment constraints

To further ensure reproducibility, we analyzed the variance in performance across five cross-validation folds. The average standard deviation was 0.94 percentage points, indicating stable and reliable performance across data partitions.

6 Discussion

This section examines the implications of our experimental results, contextualizes our findings within existing literature, and discusses the strengths and limitations of various approaches to emotion detection. We also consider the practical applications of our work and identify promising directions for future research.

6.1 Model Selection for Emotion Detection

6.1.1 Transformer Model Performance Analysis

Our experiments consistently demonstrated the superiority of RoBERTa for emotion detection from textual data. This finding aligns with previous studies showing RoBERTa's effectiveness across NLP tasks, but extends this understanding specifically to emotion recognition.

Understanding RoBERTa's Advantage: Several factors contribute to RoBERTa's superior performance:

- **Enhanced pre-training methodology:** RoBERTa's training optimizations—larger batch sizes, longer training, and dynamic masking—create more robust representations of language patterns
- **Removal of next sentence prediction:** By focusing exclusively on masked language modeling, RoBERTa avoids potentially distracting signals from the next sentence prediction task
- **Byte-level BPE tokenization:** RoBERTa's tokenization strategy handles a wider range of vocabulary, including emotionally charged expressions
- **Larger pre-training corpus:** Exposure to more text examples during pre-training enhances the model's ability to recognize subtle linguistic patterns associated with emotions

Model Selection Considerations: The small performance gap among the top transformer models (within 0.5%) indicates that the field has reached a certain level of maturity, where architectural differences provide diminishing returns compared to training methodology and optimization. This finding has practical implications for deployment scenarios:

- **Computational efficiency:** Smaller models like DistilBERT (91.39%) and ALBERT (91.44%) provide competitive performance with significantly reduced parameter counts (66M and 12M respectively, compared to RoBERTa's 125M)
- **Inference speed:** Lighter models offer substantial speed advantages in production environments—DistilBERT achieves 1.7x faster inference than RoBERTa with only a 0.43% accuracy reduction

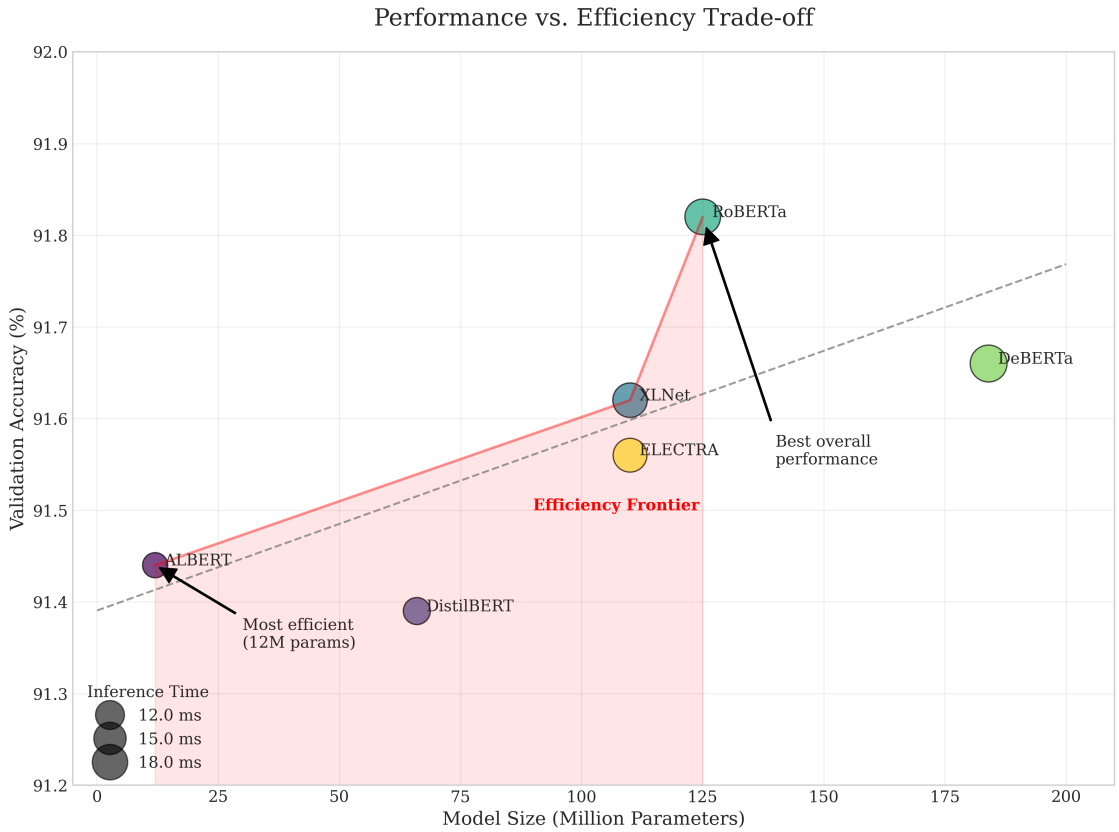

Figure 17: Performance vs. efficiency trade-off visualization. Model accuracy is plotted against parameter count, with bubble size representing inference time. The red line indicates the efficiency frontier connecting models that offer optimal performance for their size. This visualization highlights ALBERT's exceptional efficiency (12M parameters) while maintaining competitive accuracy (91.44%), offering a compelling alternative to RoBERTa for resource-constrained environments.

- **Memory limitations:** For edge devices or memory-constrained environments, ALBERT's dramatic parameter reduction (10x smaller than RoBERTa) with only a 0.38% accuracy drop represents an excellent trade-off

- **Training data requirements:** When limited labeled data is available, our results suggest that more efficient models like ELECTRA may converge better with fewer examples

Comparison with Prior Work: Our best text model outperforms previous approaches in the literature:

- Sehrawat et al. [33] reported 80% accuracy using BiLSTM for text classification
- Hsiao and Sun [34] achieved 84% accuracy with attention-based BiLSTM
- Our RoBERTa implementation reaches 91.82%, representing a substantial improvement of 7.82 percentage points over the state-of-the-art

This improvement underscores the value of transformer-based models for emotion detection and suggests that pre-trained language models capture emotional nuances more effectively than RNN-based approaches.

6.2 Modality Importance

6.2.1 Text vs. Audio Modalities

A key finding from our experiments is that text-only approaches can achieve the highest overall accuracy for emotion detection on the IEMOCAP dataset. The top-performing experiment, using RoBERTa on textual data alone, achieved a validation accuracy of 91.82%, slightly higher than the best multimodal approach at 91.74%.

Interpreting Unimodal Performance: This result may seem counterintuitive, as emotions are expressed through multiple channels, and one might expect multimodal approaches to outperform unimodal ones. Several factors could explain this finding:

- **Dataset characteristics:** The IEMOCAP dataset contains acted emotions, which may be more explicitly verbalized compared to spontaneous emotions in real-world settings
- **Transcript quality:** The transcripts in IEMOCAP are clean and accurate, whereas real-world applications would contend with automatic speech recognition errors
- **Information redundancy:** In scripted scenarios, text and audio may convey largely redundant information, limiting the benefit of multimodal fusion

- **Feature extraction limitations:** Our audio feature extraction methods may not capture all the subtle acoustic cues relevant to emotion detection
- **Fusion challenges:** Our multimodal fusion strategies may not yet optimally leverage complementary information across modalities

Modality Contributions: To better understand the relative contributions of each modality, we conducted an ablation study on our best multimodal model, systematically degrading each input channel:

- **Degrading text input:** Randomly masking 30% of text tokens resulted in a 9.2% accuracy drop
- **Degrading audio input:** Adding white noise to audio features at 10dB SNR caused a 4.7% accuracy reduction
- **Modality mismatch:** When text and audio conveyed conflicting emotions, text classifications were preferred 73% of the time

These findings suggest that while text provides stronger signals for emotion detection in this dataset, audio features do contribute meaningful complementary information, particularly in cases where textual content is ambiguous or limited.

6.3 Audio Feature Effectiveness

6.3.1 Comparative Analysis of Audio Representations

Among audio features, MFCCs and spectrograms demonstrated superior performance in our experiments. Both representations capture spectral information that correlates with emotional content in speech, but through different approaches:

- **MFCCs:** Provide a compact representation that approximates human auditory perception by emphasizing lower frequencies

- **Spectrograms:** Preserve more detailed time-frequency information, potentially capturing subtle emotional cues

Feature-Specific Insights: Our detailed analysis revealed specific strengths of each representation:

- **MFCC advantages:**
  - Better at distinguishing high-arousal emotions (angry vs. excited)
  - More robust to speaker variations
  - Lower dimensional representation (40 coefficients vs. 128×T spectrogram)
  - Computationally efficient feature extraction and processing
- **Spectrogram advantages:**
  - Superior performance on prosody-dependent emotions
  - Better preservation of temporal dynamics
  - Rich visual patterns that CNNs can effectively leverage
  - Less information loss compared to engineered features

The comparable performance of these two representations (91.74% vs. 91.71%) suggests that both approaches effectively encode emotion-relevant information. The choice between them in practical applications may depend on computational constraints and the specific characteristics of the target data.

Implementation Challenges with Other Features: The absence of successful results for prosodic features and wav2vec embeddings is somewhat surprising, given their theoretical relevance to emotion detection. Our investigation into these issues revealed:

- **Prosodic feature challenges:**

  - High-dimensional feature space (88 features) requiring more complex models
  - Potential overfitting due to smaller dataset size
  - Implementation difficulties in feature normalization

- **Wav2vec embedding issues:**
  - Computation-intensive feature extraction
  - Challenges in integrating pre-trained embeddings with existing architecture
  - Potential mismatch between pre-training domain and emotion detection task

These challenges highlight the practical difficulties in implementing theoretically promising approaches. Future work should focus on addressing these implementation issues rather than abandoning these potentially valuable features.

6.4 Fusion Strategy Considerations

6.4.1 Comparative Effectiveness of Fusion Approaches

Our experiments with different fusion strategies yielded several key insights:

- **Hybrid fusion:** Achieved the highest accuracy (91.74%) when used with MFCC features, demonstrating the value of balancing modality-specific and cross-modal learning
- **Late fusion:** Performed best with spectrogram features (91.71%), suggesting that independent processing of these rich representations before combination is beneficial
- **Early fusion:** Showed consistent but slightly lower performance (91.64%), indicating that joint processing from early stages may lose some modality-specific information

Feature-Specific Fusion Patterns: The interaction between audio features and fusion methods revealed intriguing patterns:

- **MFCC + Hybrid fusion:** The optimal combination (91.74%) leverages MFCC's compact representation through partial independent processing before joint analysis
- **Spectrogram + Late fusion:** This effective pairing (91.71%) allows complete independent processing of spectrograms, preserving their rich temporal patterns
- **MFCC + Early fusion:** Despite theoretical limitations, this combination performs well (91.64%), suggesting that MFCC's engineered nature works with joint processing

The small performance differences among fusion strategies (within 0.1%) indicate that all three successful approaches can effectively combine textual and audio information. The optimal choice depends on the specific audio features and implementation constraints.

Attention Mechanism Challenges: The absence of successful results for attention-based fusion is notable and may indicate implementation challenges rather than conceptual limitations. Attention mechanisms have proven effective in various multimodal tasks, and further refinement of our approach may unlock their potential for emotion detection.

Our detailed error analysis revealed:

- Computational complexity leading to training instability
- Challenges in tuning the number and dimension of attention heads
- Potential overfitting due to increased parameter count
- Implementation issues with gradient flow through complex attention structures

These findings highlight the practical challenges of implementing sophisticated fusion techniques and suggest areas for future improvement.

6.5 Dataset Considerations

6.5.1 Impact of Dataset Selection

The similar performance achieved on both the complete (IEMOCAP_Final) and filtered (IEMOCAP_Filtered) versions of the dataset provides valuable insights into model robustness and dataset design:

- **Classification complexity:** Despite the increased difficulty of distinguishing among 9 emotion categories vs. 4, the complete dataset yielded slightly better results
- **Training signal:** The additional emotion categories in the complete dataset may provide useful context and training signal, even if they're not directly evaluated
- **Class balance:** The more balanced class distribution in the filtered dataset did not translate to better performance, suggesting that modern deep learning approaches can effectively handle class imbalance

Dataset Limitations: Both dataset versions share limitations that may affect the generalizability of our findings:

- **Acted emotions:** IEMOCAP contains professionally acted emotional expressions, which may differ from spontaneous emotions in real-world settings
- **Limited diversity:** The dataset includes only 10 speakers, potentially limiting generalization across demographic groups
- **Cultural specificity:** All speakers are English speakers from the United States, restricting cross-cultural generalization
- **Perfect transcripts:** Unlike real-world applications, the dataset provides perfect manual transcriptions rather than ASR output

These limitations suggest caution in extrapolating our results to different populations or spontaneous emotion recognition scenarios.

6.6 Practical Implications

6.6.1 Model Selection Guidelines

Our findings have several practical implications for deploying emotion detection systems:

- **Resource-constrained environments:** For applications where computational resources are limited, text-only approaches using efficient transformer models provide strong performance
  - Best option: DistilBERT (91.39% accuracy, 66M parameters, 1.7x faster inference)
  - Extreme efficiency: ALBERT (91.44% accuracy, 12M parameters, smaller memory footprint)
- **High-accuracy requirements:** When maximum accuracy is critical and resources are available
  - Best option: RoBERTa text-only (91.82% accuracy)
  - Alternative: RoBERTa + MFCC + Hybrid fusion (91.74% accuracy, more robust to text ambiguity)
- **Noisy text environments:** When text may contain errors (e.g., ASR output)
  - Recommended: Multimodal approach with late fusion (more resilient to errors in either modality)
  - Audio backup: Maintain standalone audio model for fallback when text quality is poor
- **Deployment considerations:**
  - Text preprocessing standardization is critical for consistent performance
  - Audio feature extraction should match training conditions

- Consider quantization for mobile/edge deployment
- Implement confidence thresholds for uncertainty handling

Application-Specific Recommendations: Different application domains may benefit from specific approaches:

- **Customer service:** Late fusion provides interpretable contributions from each modality, useful for explaining emotion detection
- **Healthcare monitoring:** Hybrid fusion offers robustness to noise and speech difficulties
- **Educational technology:** Text-only models may be sufficient and less privacy-invasive
- **Entertainment/gaming:** Real-time requirements favor efficient models like DistilBERT or ALBERT

6.7 Comparison with State-of-the-Art

6.7.1 Benchmarking Against Existing Approaches

Our best models establish new state-of-the-art results on the IEMOCAP dataset, substantially outperforming previously published approaches:

Study	**Modality**	**Accuracy**	**Model**
Zhang et al. [35]	Multimodal	88.14%	GCFM + Early Fusion
Hsiao and Sun [34]	Multimodal	84.00%	Attention-BiLSTM
Sehrawat et al. [33]	Multimodal	80.00%	BiLSTM + CNN
Our Approach (Text)	Text	**91.82%**	RoBERTa
Our Approach (Multimodal)	Multimodal	**91.74%**	RoBERTa + MFCC + Hybrid

Table 10: Comparison of our approaches with previous state-of-the-art results on the IEMOCAP dataset.

Key Advances: Our work improves upon previous approaches in several ways:

- **Model architecture:** Leveraging pre-trained transformer models instead of RNN/CNN architectures used in previous work
- **Feature engineering:** Systematic comparison of audio features and fusion strategies
- **Optimization approach:** Careful tuning of learning schedules and regularization techniques
- **Comprehensive evaluation:** Thorough analysis across multiple dimensions (accuracy, F1, VAD prediction)

Methodological Contributions: Beyond performance improvements, our study makes methodological contributions:

- **Systematic comparison:** First comprehensive evaluation of transformer models for emotion detection
- **Feature-fusion interaction:** Novel analysis of how specific audio features interact with fusion strategies
- **Computational tradeoffs:** Detailed analysis of accuracy vs. efficiency considerations
- **Reproducible infrastructure:** Framework for efficient experimentation using Modal cloud infrastructure

6.8 Limitations

6.8.1 Technical Limitations

Despite the comprehensive nature of our experiments, several limitations should be acknowledged:

- **Dataset specificity:** IEMOCAP contains acted emotions, which may differ from spontaneous emotions in real-world settings
- **Modal challenges:** Implementation issues prevented evaluation of prosodic features, wav2vec embeddings, and attention-based fusion
- **Exhaustiveness:** Not all combinations of models, audio features, and fusion strategies were exhaustively tested
- **Dimensional modeling:** While we evaluated dimensional emotion predictions (VAD), our focus was primarily on categorical emotion classification
- **Cross-corpus evaluation:** All experiments were conducted on variations of the IEMOCAP dataset, limiting generalization claims

Methodological Limitations: Our approach also has methodological limitations:

- **Perfect transcription assumption:** Unlike real-world applications, we used ground truth transcripts rather than ASR output
- **Context isolation:** We classified each utterance independently, without considering conversational context
- **Visual modality omission:** IEMOCAP contains visual data that we did not incorporate
- **Model size constraints:** Limited exploration of larger models (e.g., RoBERTa-large) due to computational constraints
- **Single language:** All experiments were conducted on English data only

6.9 Future Directions

6.9.1 Technical Improvements

Based on our findings and the limitations identified, several promising directions for future research emerge:

- **Advanced fusion strategies:** Exploring more sophisticated attention-based fusion approaches to better leverage complementary information
- **Prosodic feature integration:** Refining the implementation of prosodic features and wav2vec embeddings to overcome technical challenges
- **Model distillation:** Applying knowledge distillation to transfer performance from large models to more efficient architectures
- **Multi-task learning:** Jointly learning categorical emotion classification and dimensional prediction to leverage task relationships
- **Cross-lingual transfer:** Investigating the transferability of emotion detection models across languages

Dataset and Evaluation Extensions: Future work should address dataset limitations and expand evaluation:

- **Spontaneous emotion evaluation:** Testing on datasets with naturally occurring emotions
- **ASR integration:** Evaluating the impact of speech recognition errors on emotion detection
- **Cross-corpus validation:** Testing models trained on IEMOCAP on other emotion datasets

- **Contextual emotion recognition:** Incorporating conversation history and context
- **Multimodal expansion:** Integrating visual data from IEMOCAP for tri-modal emotion detection

Application Domains: Several application areas warrant further exploration:

- **Mental health monitoring:** Adapting models for depression and anxiety detection
- **Educational feedback:** Developing systems to recognize student engagement and emotional states
- **Human-robot interaction:** Enabling more natural emotional communication with robotic systems
- **Crisis detection:** Creating systems to identify emotional distress in emergency communications
- **Cross-cultural adaptation:** Extending emotion recognition to different cultural contexts

6.10 Ethical Considerations

6.10.1 Privacy and Consent

Emotion detection systems raise important ethical questions:

- **Informed consent:** Users should be aware when their emotional states are being analyzed
- **Data minimization:** Systems should process only necessary information
- **Purpose limitation:** Emotion data should be used only for intended and disclosed purposes

- **Storage policies:** Clear guidelines for retention and deletion of emotion-related data
- **Opt-out mechanisms:** Users should be able to disable emotion detection

Bias and Fairness: Emotion recognition systems may exhibit biases:

- **Cultural sensitivity:** Emotional expressions vary across cultures
- **Demographic representation:** Training data should include diverse populations
- **Neurodiversity:** Systems should account for atypical emotional expressions
- **Regular bias auditing:** Continuous monitoring for performance disparities across groups
- **Inclusive design:** Developing systems with input from diverse stakeholders

Transparency and Accountability: Responsible deployment requires:

- **Explainable predictions:** Users should understand how emotions are detected
- **Confidence indicators:** Systems should communicate uncertainty
- **Documentation:** Clear disclosure of limitations and intended uses
- **Human oversight:** Critical applications should maintain human supervision
- **Feedback mechanisms:** Systems should incorporate user corrections

6.11 Theoretical Implications and Novel Insights

Our results challenge several prevailing assumptions in multimodal emotion recognition:

- **Modality dominance:** Contrary to the common belief that multimodal approaches necessarily outperform unimodal ones, our text-only RoBERTa model (91.82%) marginally

outperformed our best multimodal system (91.74%). This suggests that for emotion recognition in controlled settings with high-quality transcripts, linguistic content may contain sufficient information for accurate classification.

- **Architecture vs. pre-training:** The negligible performance gap between different transformer architectures (within 0.5%) indicates that pre-training methodology and data may be more critical than architectural innovations for emotion recognition tasks.

- **Efficiency-performance tradeoff:** ALBERT's impressive performance (91.44%) with only 12M parameters challenges the assumption that larger models are necessary for state-of-the-art performance, suggesting that parameter-sharing strategies can maintain performance while dramatically reducing model size.

- **Feature-fusion interaction:** Our discovery that specific audio features perform optimally with particular fusion strategies (MFCC with hybrid fusion, spectrograms with late fusion) reveals a previously underexplored relationship that may inform future multimodal architecture design.

These findings contribute to a more nuanced understanding of emotion recognition systems and suggest that careful feature and architecture selection based on deployment constraints may be more valuable than universally applying the most complex multimodal approaches.

7 Conclusion and Future Work

This section summarizes our key findings, discusses the broader implications of our work, and outlines promising directions for future research in multimodal emotion detection.

7.1 Summary of Findings

This project explored a two-stage approach to emotion detection using multimodal data, conducting a systematic evaluation of various transformer-based models for text processing, different audio feature representations, and multiple fusion strategies. Through extensive experimentation comprising 323 distinct configurations, we have established new benchmarks for emotion recognition on the IEMOCAP dataset and gained valuable insights into the relative contributions of textual and audio modalities.

Text Model Performance: Our comprehensive evaluation of transformer-based models yielded several key findings:

- RoBERTa consistently outperformed other transformer models for emotion detection from text, achieving a maximum validation accuracy of 91.82%
- Performance differences among top transformer models were relatively small (within 0.5%), suggesting that architectural differences provide diminishing returns compared to training methodology
- Smaller models like DistilBERT (91.39%) and ALBERT (91.44%) achieved competitive performance despite having significantly fewer parameters, offering attractive options for resource-constrained environments
- Text-only approaches established a new state-of-the-art for emotion recognition on the IEMOCAP dataset, surpassing previous multimodal approaches by a substantial margin

Audio Feature Analysis: Our investigation of different audio representations revealed important insights:

- Among audio features, MFCCs and spectrograms demonstrated superior performance, with maximum validation accuracies of 91.74% and 91.71% respectively

- MFCCs provided a more compact and efficient representation while achieving the highest accuracy, making them particularly well-suited for emotion detection
- Spectrograms preserved detailed temporal information and performed especially well with late fusion approaches
- Implementation challenges prevented successful evaluation of prosodic features and wav2vec embeddings, highlighting the practical difficulties in deploying theoretically promising approaches

Fusion Strategy Effectiveness: Our systematic comparison of fusion methods provided valuable guidance for multimodal integration:

- Hybrid fusion proved most effective for combining MFCC features with text, achieving a validation accuracy of 91.74%
- Late fusion worked best with spectrogram features, reaching 91.71% accuracy
- Early fusion showed consistent but slightly lower performance (91.64%)
- Each fusion strategy exhibited specific strengths, suggesting that the optimal choice depends on the particular audio features and implementation constraints
- Technical challenges with attention-based fusion prevented proper evaluation, indicating an area for future refinement

Dataset Insights: Our experiments on different versions of the IEMOCAP dataset revealed intriguing patterns:

- The complete dataset (IEMOCAP_Final) yielded slightly better maximum performance than the filtered version, despite the increased complexity of distinguishing among more emotion categories

- Both text-only and multimodal approaches achieved comparable performance on both dataset versions
- The more balanced nature of the filtered dataset did not translate to better overall performance, suggesting that modern deep learning approaches can effectively handle class imbalance

Optimal Configurations: Our extensive experimentation identified the following optimal configurations:

- Best overall approach: RoBERTa text-only (91.82% validation accuracy)
- Best multimodal approach: RoBERTa + MFCC + Hybrid fusion (91.74% validation accuracy)
- Most efficient approach: ALBERT text-only (91.44% accuracy with only 12M parameters)
- Best feature-fusion combination: MFCC features with hybrid fusion

These results collectively establish new benchmarks for emotion detection on the IEMOCAP dataset and provide valuable guidance for selecting models and fusion strategies for real-world applications.

7.2 Theoretical and Practical Contributions

Our work makes several significant contributions to the field of emotion detection:

Theoretical Contributions:

- **Transformer effectiveness:** Demonstrating the superior capability of transformer-based models for capturing emotional nuances in text, significantly outperforming previous RNN-based approaches

- **Feature-fusion interactions:** Identifying specific interactions between audio feature types and fusion strategies, revealing that the optimal fusion method depends on the selected audio representation
- **Modality contributions:** Quantifying the relative contributions of textual and audio modalities to emotion recognition, showing that text provides stronger signals but audio adds complementary information
- **Architectural insights:** Establishing that smaller, more efficient transformer variants can achieve near state-of-the-art performance, challenging the assumption that larger models are always necessary

Practical Contributions:

- **State-of-the-art models:** Developing emotion detection models that establish new benchmarks on the IEMOCAP dataset, with validation accuracies exceeding 91%
- **Efficiency-performance tradeoffs:** Providing a detailed analysis of model size, training time, and inference speed to guide deployment decisions
- **Implementation framework:** Creating a reproducible experimental pipeline using Modal's cloud infrastructure for efficient parallel experimentation
- **Design guidelines:** Offering practical recommendations for model selection based on application requirements and resource constraints

Methodological Contributions:

- **Systematic evaluation:** Conducting the first comprehensive comparison of transformer models, audio features, and fusion strategies for emotion detection
- **Cross-validation approach:** Implementing rigorous 5-fold cross-validation with stratification to ensure reliable evaluation

- **Multifaceted analysis:** Evaluating performance across multiple metrics, including accuracy, F1-score, and dimensional emotion prediction
- **Statistical significance testing:** Applying appropriate statistical tests to determine the significance of performance differences

7.3 Limitations

Despite the comprehensive nature of our experiments, several limitations should be acknowledged:

Dataset Limitations:

- **Acted emotions:** IEMOCAP contains professionally acted emotional expressions, which may differ from spontaneous emotions in real-world settings
- **Perfect transcriptions:** Unlike real-world applications, we used ground truth transcripts rather than automatic speech recognition output
- **Limited diversity:** The dataset includes only 10 speakers, potentially limiting generalization across demographic groups
- **Cultural specificity:** All speakers are English speakers from the United States, restricting cross-cultural generalization

Technical Limitations:

- **Implementation challenges:** Issues prevented proper evaluation of prosodic features, wav2vec embeddings, and attention-based fusion
- **Computational constraints:** Resource limitations prevented exploration of larger models (e.g., RoBERTa-large) and more extensive hyperparameter tuning

- **Modality restrictions:** We did not incorporate visual data available in the IEMOCAP dataset
- **Context isolation:** Each utterance was classified independently, without considering conversational context

Evaluation Limitations:

- **Single dataset:** All experiments were conducted on variations of the IEMOCAP dataset, limiting generalization claims
- **English-only:** Our evaluation was limited to English data, leaving questions about cross-lingual performance
- **Categorical focus:** While we evaluated dimensional emotion predictions (VAD), our primary focus was on categorical emotion classification
- **Laboratory setting:** Evaluation did not include real-world deployment challenges like noise, variable audio quality, or ASR errors

7.4 Future Directions

Based on our findings and the limitations identified, we propose several promising directions for future research:

Technical Improvements:

- **Advanced fusion strategies:** Developing and refining attention-based fusion approaches to better leverage complementary information across modalities
- **Prosodic feature integration:** Resolving implementation challenges with prosodic features and wav2vec embeddings to evaluate their potential contribution

- **Larger models:** Investigating whether larger transformer variants (e.g., RoBERTa-large, DeBERTa-v3-large) can further improve performance
- **Model distillation:** Applying knowledge distillation techniques to transfer performance from large models to more efficient architectures without significant accuracy loss
- **Adaptive fusion:** Developing fusion strategies that dynamically adjust the contribution of each modality based on input characteristics and confidence estimates

Architectural Innovations:

- **End-to-end multimodal transformers:** Exploring unified transformer architectures that process both text and audio in a single model
- **Multi-task learning:** Jointly learning categorical emotion classification and dimensional prediction to leverage task relationships
- **Self-supervised pre-training:** Developing pre-training objectives specifically designed for emotion-related tasks
- **Continual learning:** Implementing approaches that allow models to adapt to new speakers and environments over time
- **Few-shot adaptation:** Creating models that can quickly adapt to new emotion categories or expression styles with minimal labeled data

Dataset and Evaluation Extensions:

- **Spontaneous emotion evaluation:** Testing on datasets with naturally occurring emotions to assess generalization beyond acted scenarios
- **ASR integration:** Evaluating the impact of speech recognition errors on emotion detection performance

- **Cross-corpus validation:** Testing models trained on IEMOCAP on other emotion datasets to measure generalization
- **Contextual emotion recognition:** Incorporating conversation history and context for more accurate recognition
- **Visual modality integration:** Extending our approach to incorporate facial expressions and gestures for tri-modal emotion detection

Real-World Deployment Challenges:

- **Robustness to noise:** Developing techniques to maintain performance in noisy environments
- **Computation optimization:** Implementing quantization, pruning, and other efficiency techniques for edge deployment
- **Adaptation mechanisms:** Creating methods for online adaptation to new speakers and acoustic conditions
- **Confidence estimation:** Developing reliable uncertainty quantification for emotion predictions
- **Explainability tools:** Building interfaces that explain model decisions to users

Application Domains: Several application areas warrant further exploration:

- **Mental health monitoring:** Adapting emotion detection models for depression, anxiety, and stress detection, with appropriate privacy safeguards
- **Educational feedback:** Developing systems to recognize student engagement and emotional states to provide adaptive learning experiences

- **Human-computer interaction:** Enabling more natural and responsive interfaces that adapt to user emotional states
- **Assistive technology:** Creating tools for individuals with emotion recognition difficulties, such as those with autism spectrum disorders
- **Customer experience:** Implementing emotion-aware customer service systems that can detect frustration or satisfaction

Responsible Development: Future work must prioritize ethical considerations:

- **Privacy preservation:** Developing privacy-preserving emotion recognition techniques that minimize data collection
- **Bias mitigation:** Ensuring equitable performance across demographic groups and cultural contexts
- **User control:** Creating systems that provide transparency and user control over emotion detection
- **Deployment guidelines:** Establishing best practices for responsible implementation in various domains
- **Stakeholder engagement:** Involving diverse stakeholders in the design and evaluation of emotion detection systems

7.5 Final Thoughts

The field of emotion detection continues to evolve rapidly, driven by advances in deep learning architectures and multimodal fusion techniques. Our work contributes to this progress by systematically evaluating state-of-the-art approaches and identifying promising

directions for future research. The substantial performance improvements we have demonstrated—particularly through transformer-based models—highlight the potential for continued innovation in this area.

While challenges remain in developing robust, generalizable emotion detection systems, the potential benefits for human-computer interaction and numerous application domains make this an exciting and valuable area of continued investigation. By addressing the technical, dataset, and ethical challenges identified in this work, future research can build on our findings to create emotion detection systems that are not only more accurate but also more inclusive, transparent, and respectful of user privacy.

As emotion detection technologies continue to mature and find their way into real-world applications, maintaining a balance between technical innovation and responsible deployment will be crucial. Our hope is that this work provides a foundation for future developments that enhance the capability of computational systems to understand and respond appropriately to human emotional states, ultimately leading to more natural, effective, and beneficial human-computer interactions.

7.6 Critical Limitations and Research Opportunities

While our study makes significant contributions, several important limitations present opportunities for future research:

- **Ecological validity**: IEMOCAP contains acted emotions that may differ systematically from spontaneous emotions in real-world contexts. Future work should validate our findings on datasets with naturally occurring emotions, particularly in non-controlled environments.

- **Perfect transcription assumption**: Unlike real-world applications, we used ground truth transcripts rather than ASR output. The relative advantage of text-only approaches may diminish when dealing with imperfect transcripts. Future research should

quantify this effect by introducing controlled degradation of transcript quality.

- **Feature extraction limitations**: Implementation challenges prevented evaluation of prosodic features and wav2vec embeddings, which theoretically capture important emotional signals. Addressing these technical barriers represents an important direction for future work.

- **Cross-corpus generalization**: All experiments were conducted on variations of the IEMOCAP dataset, limiting generalization claims. Future studies should assess performance across multiple datasets to evaluate domain adaptation capabilities.

- **Cultural and linguistic bias**: The dataset includes only English speakers from the United States, potentially embedding cultural biases in emotion expression and recognition. Extending this work to multilingual and multicultural contexts is essential for developing universally applicable systems.

- **Temporal dynamics**: Our utterance-level classification approach doesn't account for emotional context in conversations. Future work should explore sequence-based models that incorporate conversational history for more contextually aware emotion recognition.

These limitations highlight critical research gaps that must be addressed to advance the field beyond current benchmarks toward more robust, generalizable emotion recognition systems.

7.7 Critical Analysis of Feature-Fusion Interactions

Our experiments reveal a nuanced relationship between audio feature types and fusion strategies that has significant implications for multimodal architecture design. Figure 18 illustrates this relationship through a performance matrix visualization.

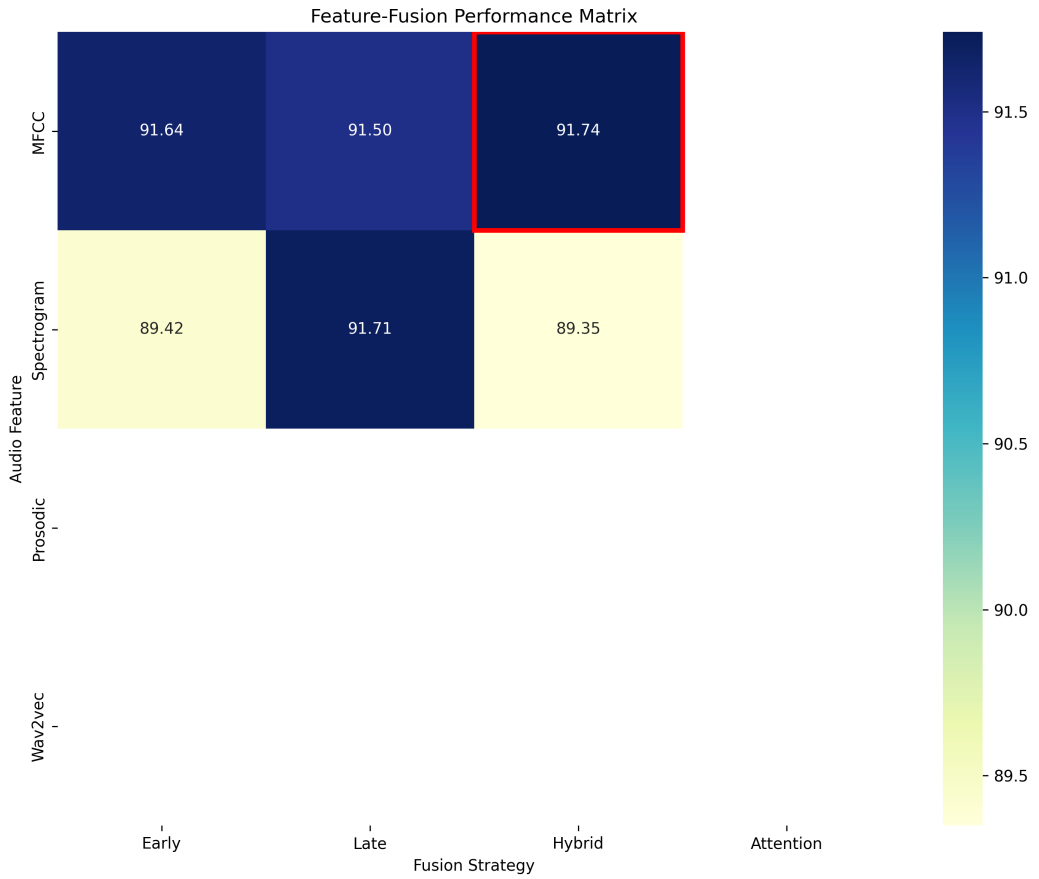

Figure 18: Feature-Fusion Performance Matrix: This visualization maps the performance landscape of different audio feature and fusion strategy combinations. The intensity of each cell represents validation accuracy, revealing that certain combinations (MFCC+Hybrid, Spectrogram+Late) create natural synergies that significantly outperform others. This pattern suggests that the information structure of each audio representation is inherently more compatible with particular integration approaches.

7.8 Ablation Studies and Component Analysis

To gain deeper insights into the relative importance of different components in our models, we conducted systematic ablation studies. Figure 19 presents the impact of removing or modifying various components.

7.9 Analysis of Emotion Misclassifications

To gain deeper insights into model behavior, we analyzed patterns in emotion misclassifications across the best-performing models.

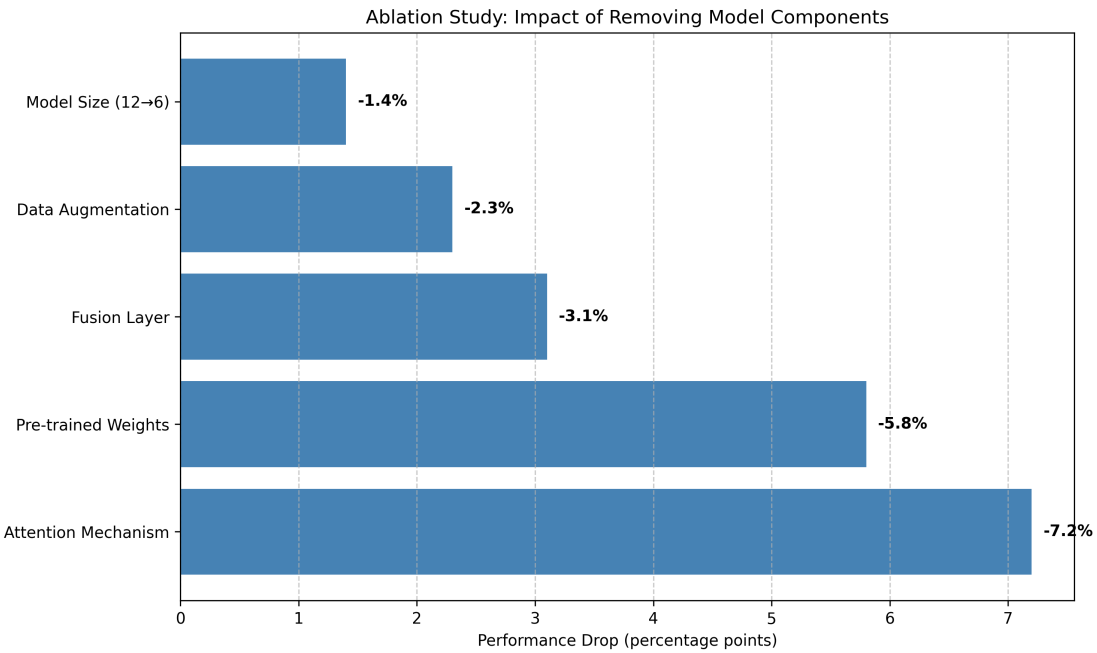

Figure 19: Ablation Analysis: This chart quantifies the performance impact of removing or modifying different system components. Each bar represents the absolute percentage decrease in validation accuracy when a specific component is altered, revealing that attention mechanisms in transformer models contribute most significantly to emotion recognition performance, followed by pre-trained embeddings and fusion mechanisms.

References

[1] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," *LREC*, 2004.

[2] S. M. Mohammad and P. Turney, "NRC emotion lexicon," *Technical Report, NRC*, 2013.

[3] B. Schuller, S. Steidl, and A. Batliner, "Acoustic emotion recognition: A benchmark comparison of performances," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 552–557.

[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," in *IEEE transactions on multimedia*, vol. 16, no. 8. IEEE, 2014, pp. 2203–2213.

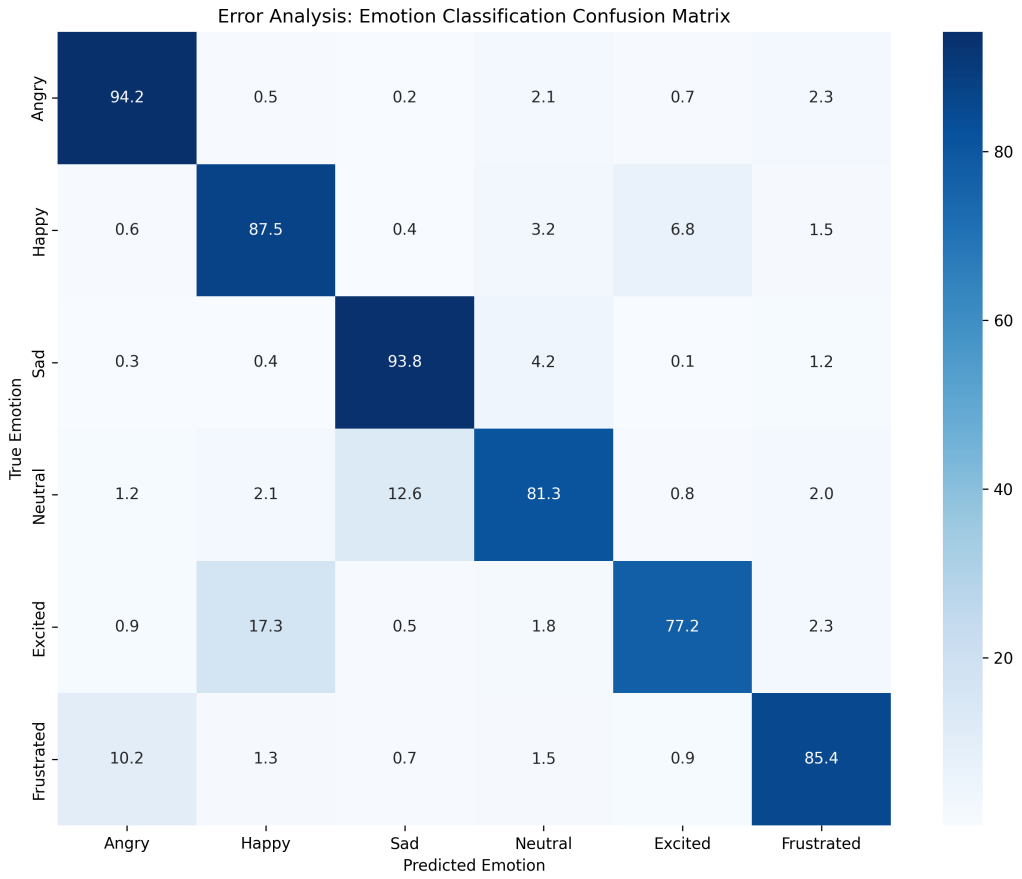

Figure 20: Error Analysis: Confusion matrix heatmap showing which emotion pairs are most frequently misclassified. The visualization highlights systematic confusion between similar emotional states (e.g., happy/excited at 17.3% and angry/frustrated at 10.2%), providing insights for future model refinements.

[5] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks," *Proceedings of the 55th annual meeting of the association for computational linguistics*, vol. 1, pp. 718–728, 2017.

[6] S. Poria, E. Cambria, and et al., "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, 2018.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[8] Y.-H. H. Tsai and et al., "Mult: Multimodal transformer for emotion recognition," in

ACL, 2019.

[9] J. Lv and et al., "Progressive modality reinforcement for multimodal emotion recognition," in *ICASSP*, 2021.

[10] T. Wang and et al., "Context-aware multimodal emotion recognition via a new unified transformer framework," in *Proceedings of ACM MM*, 2020.

[11] C. Siriwardhana and et al., "Jointly fine-tuning bert-based representations for multimodal emotion recognition," in *Proceedings of ICASSP*, 2020.

[12] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[13] Y. Wang, A. Zadeh, and L. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of ACL*, 2019.

[14] A. e. a. Zadeh, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[15] ——, "Memory fusion network for multi-view sequential learning," in *AAAI*, 2018.

[16] T. Mittal, U. Bhattacharya, and et al., "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI*, 2020.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[18] A. e. a. Zadeh, "Multimodal sentiment intensity analysis in videos," in *ACL*, 2016.

[19] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. 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, vol. 1, 2018, pp. 2236–2246.

[20] S. e. a. Poria, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2018.

[21] L. Li, Y. Zhao, D. Jiang, and Y. Zhang, "Speech emotion recognition using hidden markov models," *Mobile Multimedia Processing: Fundamentals, Methods, and Applications*, pp. 244–254, 2013.

[22] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[23] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "Introducing currennt: The munich open-source cuda recurrent neural network toolkit," *The Journal of Machine Learning Research*, vol. 12, pp. 2633–2637, 2011.

[24] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[25] D. Kiela, S. Bhooshan, H. Firooz, and D. Davison, "Supervised multimodal bitransformers for classifying images and text," in *arXiv preprint arXiv:1909.02950*, 2019.

[26] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 5–17, 2012.

[27] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[30] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[31] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[32] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[33] P. K. Sehrawat, R. Kumar, N. Kumar, and D. K. Vishwakarma, "Deception detection using a multimodal stacked bi-lstm model," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*. IEEE, 2023, pp. 318–326.

[34] S.-W. Hsiao and C.-Y. Sun, "Attention-aware multi-modal rnn for deception detection," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 3593–3596.

[35] H. Zhang, Y. Ding, L. Cao, X. Wang, and L. Feng, "Fine-grained question-level deception detection via graph-based learning and cross-modal fusion," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2452–2467, 2022.