Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.086256

ARTCILE

Multimodal Emotion Recognition in Urdu through Late Fusion of Fine-Tuned Speech and Text Representations

Muhammad Sheraz[1], Adil Majeed[1], Shehzad Khalid[2,3,*], Yazeed Alkhrijah[4,*], Sulieman S. Alshuhri[5] and Hasan Mujtaba[1]

[1]Department of Artificial Intelligence and Data Science, National University of Computer & Emerging Sciences, AK Brohi Rd., H-11/4, Islamabad, Pakistan

[2]Computer and Information Sciences Research Center (CISRC), Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

[3]Department of Computer Engineering, Bahria School of Engineering and Applied Sciences (BSEAS), Bahria University, Islamabad, Pakistan

[4]Department of Electrical Engineering, Imam Mohammad ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

[5]Department of Information Technology, College of Computer and Information Sciences, Imam Mohammad ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

*Corresponding Authors: Shehzad Khalid. Email: shehzad@bahria.edu.pk; Yazeed Alkhrijah. Email: Ymalkhrijah@imamu.edu.sa

Received: 27 May 2026; Accepted: 29 July 2026; Published: 28 August 2026

ABSTRACT: Emotion recognition plays a crucial role in enabling intelligent human–computer interaction, yet research in low-resource languages such as Urdu remains limited, particularly in multimodal settings. This study proposes a multimodal deep learning framework for Urdu emotion recognition by integrating speech and text modalities. The approach leverages transformer-based models, namely wav2vec 2.0 for audio representation and MuRIL for text representation, combined using a late fusion strategy for classification. Experiments were conducted on the UMED dataset, consisting of 8269 multimodal instances across five emotion classes. The proposed multimodal model achieved an accuracy of 0.701 and an F1-score of 0.6915, outperforming unimodal baselines, where the audio-only and text-only models achieved accuracies of 0.6681 and 0.5085, respectively. Furthermore, the proposed approach surpasses the existing UMEDNet benchmark, demonstrating the effectiveness of transformer-based feature extraction and multimodal late fusion for Urdu emotion recognition. The results highlight the complementary nature of speech and text modalities and demonstrate that independently learned modality-specific classifiers combined through decision-level fusion can improve emotion recognition performance in low-resource languages. However, the performance improvement over alternative fusion strategies was relatively modest, indicating that more advanced multimodal interaction mechanisms may further enhance recognition performance.

KEYWORDS: Multimodal; emotion detection; audio emotion detection; text emotion detection

1 Introduction

Emotions play a crucial role in communication, decision-making, and overall well-being, influencing both personal relationships and everyday actions [1]. Developing systems capable of automatically recognizing human emotions remains a challenging task because emotional expressions are highly complex and can be conveyed through multiple modalities, including speech, text, facial expressions, and body language. Despite these challenges, emotion recognition has become an important research area in artificial intelligence (AI) because emotions significantly influence human behavior, decision-making, and social interactions [2].

Applications of automatic emotion recognition have expanded considerably with recent advances in AI [3]. Emotion recognition can serve as a valuable, non-invasive tool for the early detection and monitoring of mental health conditions such as depression and anxiety [4]. In customer service, emotion recognition systems can identify emotions such as frustration, happiness, and anger, enabling organizations to assess customer sentiment and improve service quality [3]. In call centers, these systems can detect customer agitation or satisfaction, prioritize urgent calls, and route them to appropriate agents [5]. Emotion recognition also has important applications in education, where it can help instructors better understand students' emotional states and create more responsive learning environments [6]. Similarly, it enhances user experience in video games by enabling more immersive and personalized interactions [7]. Additional applications include public safety, where surveillance systems can identify signs of fear or distress [8], and intelligent transportation systems, where monitoring drivers' emotional states such as anger and stress may help reduce accident risks [9]. These diverse applications highlight the importance of developing accurate and reliable emotion recognition systems.

Human emotions are naturally expressed through multiple modalities, including speech, text, facial expressions, and physiological signals. Consequently, emotion recognition has traditionally been studied using individual modalities [9,10]. However, recent advances in AI have shifted research toward multimodal emotion recognition, where information from multiple sources is integrated to improve prediction performance [11,12]. In multimodal emotion recognition, different modalities provide complementary information. Speech captures acoustic and prosodic characteristics such as pitch, intensity, and speaking rate, whereas text provides semantic and contextual information. Combining these complementary sources enables multimodal systems to learn richer emotional representations than unimodal approaches. Various fusion strategies have been proposed for multimodal integration, including feature-level fusion and decision-level late fusion.

Although multimodal emotion recognition has received considerable attention in recent years, most existing studies focus on resource-rich languages such as English [11] and French [13]. This is largely due to the availability of large multimodal datasets and well-developed pre-trained language and speech models for these languages. In contrast, Urdu remains a low-resource language with limited publicly available multimodal datasets and pre-trained resources, despite being one of the most widely spoken languages in the world, with an estimated 231.7 million speakers. While several studies have investigated unimodal emotion recognition in Urdu [14,15], to the best of our knowledge, only one multimodal study has been reported [12]. That work incorporates speech, text, and video modalities. However, no previous study has specifically investigated multimodal emotion recognition using only speech and text modalities for Urdu. This gap highlights the need for further research on efficient multimodal emotion recognition frameworks for low-resource languages.

This paper makes the following contributions:

1. Design and develop an Urdu-specific multimodal emotion recognition framework that integrates fine-tuned speech and text representations through a decision-level late fusion strategy.
2. Implement unimodal baselines (speech-only and text-only) and systematically compare them with the proposed multimodal system.
3. Evaluate the effectiveness of multimodal fusion in improving classification accuracy and F1-score for Urdu, a resource-poor language.
4. Provide comprehensive experimental results and baseline comparisons that can serve as a reference for future research on multimodal emotion recognition in Urdu.

Rather than introducing a new multimodal fusion algorithm, this work demonstrates how state-of-the-art pretrained speech and language models can be effectively adapted and integrated for multimodal emotion recognition in the low-resource Urdu language.

The rest of the paper is organized as follows: The related work section discusses the relevant literature, followed by the proposed methodology and experimentation section.

2 Literature Review

In recent years, many researchers have published their works on emotion recognition utilizing speech and text modality as well as the combination of both. This section addresses several of these studies conducted in various languages, including English, Chinese, and French. However, according to the literature, there is very little work published for Urdu in the multimodal domain.

2.1 Speech Modality in Emotion Recognition

The authors [16] proposed a Multilayer Perceptron (MLP)–based neural network for speech emotion recognition, using Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction. The system classified seven emotions, i.e., worry, surprise, neutral, sadness, happiness, hate, and love, and was evaluated on RAVDESS, TIMIT, and Emo-DB datasets. The model achieved accuracies of 81.02%, 84.23%, and 86.71%, respectively. The researchers [17] proposed a hybrid feature-based Speech Emotion Recognition (SER) technique that combines Mel Frequency Cepstral Coefficients (MFCCs) with time-domain (t-domain) features collectively called MFCCT features and used a 1D Convolutional Neural Network (CNN) for classification. The approach was evaluated on benchmark datasets: SAVEE and RAVDESS. The proposed method achieved accuracies of 92.6% and 91.4%, respectively. The authors proposed an attention-based deep learning model for speech emotion recognition, combining a two-dimensional Convolutional Neural Network (CNN-2D) with Long Short-Term Memory (LSTM) layers and a self-attention mechanism [18]. The system used Mel Frequency Cepstral Coefficients (MFCCs) as the primary features. It was trained on a custom dataset combining RAVDESS, SAVEE, and TESS corpora to detect eight emotions. The proposed CNN–LSTM–Attention model achieved a 90.19% average accuracy.

2.2 Speech Modality in Urdu Emotion Recognition

The authors in [19] proposed a Speech Emotion Recognition (SER) approach for Pakistani Urdu using the URDU dataset and a Bidirectional Gated Recurrent Unit (Bi-GRU) model. They extract key acoustic features, including MFCCs, GFCCs, and eGeMAPS, and employ SHAP analysis for model interpretability. Their method achieves a validation accuracy of approximately 91%. The authors [15] explore emotion recognition in Urdu speech using Uni-GRU and Bi-GRU architectures applied to Mel-spectrogram features. Using the URDU dataset, they first implement a Uni-GRU model, achieving 81.25% accuracy. Researchers [20] apply a deep learning approach based on a 1D-CNN architecture, leveraging spectral speech features for emotion classification. The proposed model achieved an accuracy of 97% on the URDU dataset. The authors [21] present a novel Speech Emotion Recognition (SER) framework for Urdu, combining MFCC features with CNN-GRU architectures, using the SEMOUR+ dataset, along with cross-validation on an additional Urdu dataset. Their approach achieves a maximum accuracy of 84.92%.

2.3 Text Modality in Emotion Recognition

They proposed a textual emotion recognition model that integrates the ALBERT-BiLSTM architecture with a hybrid SVM-NB (Support Vector Machine–Naive Bayes) classifier to enhance the accuracy of text-based emotion detection [22]. Experiments on three Chinese datasets, namely, ChineseNlpCorpus,

NLP&CC 2014, and WEC, showed that the proposed model achieved 92.95%, 91.26%, and 93.11% accuracy, respectively. They explore a transformer-based approach for Bengali text emotion classification to address challenges in low-resource languages [23]. A new dataset named BEmoC, containing 6243 manually annotated texts across six emotions was developed for the task. The study compared multiple models, including traditional ML, deep neural networks, and transformers. Among all, the XLM-R model achieved the best performance with 69.73%. The authors in the research [24] proposed a hybrid deep learning model combining Convolutional Neural Networks (CNN) and Bidirectional Gated Recurrent Units (Bi-GRU) enhanced with pre-trained Neural Network Language Model (NNLM) embeddings for emotion detection in conversational text. The model leverages CNN layers to extract local contextual features and Bi-GRU layers to capture long-range dependencies. Among all evaluated architectures, the CNN-BiGRU with NNLM embeddings achieved the best performance, with a testing accuracy of 73.74%.

2.4 Text Modality in Urdu Emotion Recognition

Authors [25] address the lack of resources for emotion detection in Roman Urdu text by creating a new 18k sentence annotated corpus labeled across six emotion classes. To classify emotions, they propose Deep-EmoRU, a hybrid deep learning model combining LSTM and CNN feature learners. The proposed model outperforms all baselines, achieving an accuracy of 82.2%. The authors [26] proposed a CNN–LSTM-based framework for emotion detection and sentiment analysis in the low-resource Urdu language. Their approach combines convolutional layers for local feature extraction with LSTM layers to capture long-range contextual dependencies. Evaluated on Urdu text data, the proposed method achieved an emotion detection accuracy of 95%, demonstrating the effectiveness of deep learning architectures for Urdu text emotion analysis. Authors [14] introduced the Urdu Nastalique Emotions Dataset (UNED), a publicly available corpus annotated with six emotions in both sentences and paragraphs. They propose a deep learning-based model for emotion classification on this low-resource Urdu dataset. Through extensive experiments, the deep learning model outperforms traditional machine learning methods, achieving an F1 score of 85%. Ref. [27] proposes a novel approach for emotion detection in Roman Urdu text using the XLM-R (Roberta) transformer model. Their method classifies text into six emotional categories. The model outperforms traditional machine learning techniques, achieving an F1-score of 85% and 84% accuracy.

2.5 Speech and Text Modality in Emotion Recognition

Researchers [28] proposed a multimodal deep learning approach for emotion recognition by fusing acoustic and text data. Acoustic features are extracted using a SincNet layer with DCNN, while text is processed through parallel DCNN and Bi-RNN+DCNN branches with cross-attention. Evaluated on the IEMOCAP dataset, achieved a 5.2% improvement in weighted accuracy over state-of-the-art baselines. Authors proposed a method that fuses speech and text by introducing a two-layer Dynamic Bayesian Mixture Model (2L-DBMM) for multimodal fusion [29]. On EmoUERJ (Portuguese) and ESD (English) datasets, it achieved up to 98% accuracy. Researcher [30] introduces a bimodal speech emotion recognition system that fuses acoustic features and linguistic information through a decision-level strategy combining both emotion and sentiment cues. Using the RAMAS database, the bimodal system outperformed unimodal approaches, achieving a UAR of 72.01%. Authors [13] fine-tuned pre-trained Transformer encoders for speech and text emotion recognition on the CEMO French emergency call center dataset, and evaluated different fusion strategies. Results show that multimodal fusion improved performance by 4%–9% over single modalities.

2.6 Speech and Text Modality in Urdu Emotion Recognition

The authors in [12] proposed UMEDNet (Urdu Multimodal Emotion Detection Network), which, to the best of our knowledge, is the only published multimodal emotion recognition framework specifically developed for the Urdu language at the time of writing. The model incorporates three modalities, i.e., facial frames, audio, and text for classification. They utilized MTCNN and FaceNet for visual feature extraction. For audio features, they fine-tuned wav2vec to extract Urdu Speech Embeddings and XLM-R Roberta for text processing. These features were concatenated for classification. Even though the model deals with three modalities, the authors also developed a model using speech and text features only. This model was trained over the UMED dataset, which consists of about 9000 data examples belonging to five classes, namely happiness, sadness, anger, love, and neutral. The speech and text combined model achieved an accuracy of 62%, a precision of 60%, a recall of 62%, and an F1-score of 61%.

The summary of the related work is shown in Table 1.

Table 1: Summary of related work in emotion recognition.

Reference	Year	Language	Modality	Methodology	Evaluation Metrics
[16]	2022	English	Speech	MLP	Acc 81% (RAVDESS), Acc 84% (TIMIT), Acc 86% (EmoDB)
[17]	2023	English	Speech	CNN	Acc 92% (SAVEE), Acc 91% (RAVDESS)
[18]	2023	English	Speech	LSTM+CNN	Acc 90.19% (TESS+RAVDESS+SAVEE)
[19]	2024	Urdu	Speech	Bi-GRU Network	Acc 91%
[15]	2025	Urdu	Speech	Uni-GRU, Bi-GRU	Acc 81.25%, 87.50%
[20]	2023	Urdu	Speech	CNN	Acc 97%
[21]	2023	Urdu	Speech	CNN+GRU	Acc 84.92%
[22]	2023	Chinese	Text	BiLSTM+SVM	Acc 92% (ChineseCorpus), Acc 91% (NLP&CC), Acc 93% (WEED)
[23]	2021	Bengali	Text	XLM-R	Acc 69% (BEmoC)
[24]	2024	English	Text	Bi-GRU+CNN	Acc 73% (EDD)
[25]	2022	Roman Urdu	Text	CNN+LSTM	Acc 82.2%, F-score 0.82
[26]	2022	Urdu	Text	CNN–LSTM	Accuracy 95.00%
[14]	2023	Urdu	Text	DL	F1 85% (sentences), F1 50% (paragraphs)
[27]	2025	Roman Urdu	Text	XLM-R (Roberta)	Acc 84%, F1 85%
[28]	2020	English	Speech+Text	Bi-RNN+DCNN	Acc 79%
[29]	2024	Portuguese	Speech+Text	2L-DBMM	Acc 98%
[30]	2021	Russian	Speech+Text	–	Acc 72.01%
[13]	2023	French	Speech+Text	Transformer Encoder	4%–9% better accuracy than unimodal
[12]	2025	Urdu	Speech+Text	XLM-R Roberta + Wav2vec	Acc 62%, F1-score 61%, Precision 60%, Recall 62%

Despite the considerable progress reported in the literature, several important research gaps remain. Existing speech emotion recognition studies, particularly for Urdu, largely rely on handcrafted acoustic features or conventional deep learning architectures such as CNNs and recurrent neural networks. Likewise, many text-based approaches focus on unimodal emotion recognition and do not investigate how textual information can complement acoustic cues. Although transformer-based models have recently demonstrated superior representation learning capabilities, their application to multimodal Urdu emotion recognition remains limited. Furthermore, most multimodal studies have been conducted for resource-rich languages, while only one prior study has explored multimodal emotion recognition in Urdu using audio, text, and visual modalities. To the best of our knowledge, no previous work has systematically investigated a transformer-based speech–text framework using independently fine-tuned modality-specific models combined through a late fusion strategy. These limitations motivate the proposed approach, which aims to exploit complementary speech and textual representations while providing a comprehensive comparison with unimodal baselines and alternative fusion strategies.

Overall, the literature demonstrates that recent advances in deep learning and transformer-based models have significantly improved emotion recognition performance across different modalities. However, most existing studies either focus on unimodal emotion recognition or target resource-rich languages where large datasets and pretrained models are readily available. For Urdu, multimodal emotion recognition remains largely unexplored, with only one previously reported multimodal framework. Moreover, limited attention has been given to investigating transformer-based speech and text representations within a decision-level late fusion framework. These observations motivate the proposed study, which develops and evaluates a multimodal speech–text emotion recognition framework specifically designed for the Urdu language.

3 Methodology

This study proposes a multimodal deep learning framework for Urdu emotion recognition by integrating speech and text information. The overall pipeline consists of four stages: dataset preparation, unimodal representation learning, embedding extraction, and multimodal late fusion classification. The motivation behind this design is that emotion is rarely expressed through a single channel alone. In spoken language, emotional meaning is conveyed not only through lexical content but also through vocal characteristics such as pitch, speaking rate, stress, and prosody. A multimodal framework provides a more complete representation of emotional expression than either speech-only or text-only systems.

The proposed framework first fine-tunes transformer-based models independently for the speech and text modalities to learn modality-specific emotional representations. Speech embeddings are extracted using a fine-tuned wav2vec 2.0 model [31], while textual representations are generated using a fine-tuned MuRIL [32] based Multiple Instance Learning framework. These embeddings are then processed through separate modality-specific classification branches, and final emotion prediction is obtained using a late fusion strategy based on decision-level averaging of classifier outputs. Fig. 1 gives an overview of the proposed system.

3.1 Audio Representation Learning

Emotion recognition from speech has traditionally relied on handcrafted acoustic features such as MFCCs, pitch statistics, energy, and spectrogram-based descriptors. While such features can be useful, they are often limited by the assumptions built into the feature design process. In contrast, modern self-supervised speech models learn rich latent representations directly from raw waveform data and have shown strong transfer performance across a wide range of downstream tasks. To capture emotionally relevant acoustic

cues, this work employed the pretrained wav2vec2 model as the base audio encoder. This model is derived from the Wav2Vec2 family, which is a transformer-based speech representation framework designed to learn contextualized features from raw audio as shown in Algorithm 1. A key advantage of Wav2Vec2 is that it eliminates the need for manual acoustic feature engineering by learning task-relevant representations automatically. Although the selected XLS-R-300M model was pretrained on multilingual speech data rather than specifically on Urdu, its cross-lingual speech representations make it suitable for modeling Urdu speech patterns.

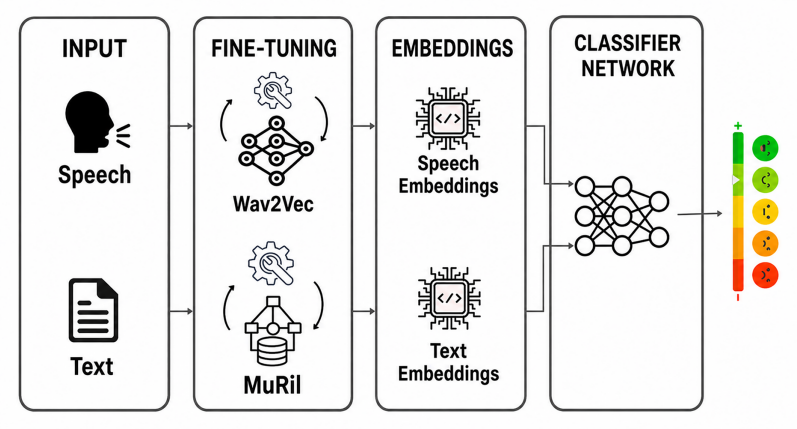

Figure 1: Proposed system architecture. The proposed multimodal emotion recognition framework, where speech and text embeddings are extracted using fine-tuned Wav2Vec and MuRIL models and combined for final emotion prediction.

Algorithm 1: Fine-tuning Wav2Vec2 for Urdu speech emotion recognition

Require: Training dataset $\mathcal{D}_{train}$, validation dataset $\mathcal{D}_{val}$, pretrained Wav2Vec2 model M
Ensure: Fine-tuned speech emotion recognition model
Initialize pretrained Wav2Vec2 model with emotion classification head
Freeze low-level feature extraction layers
Compute class weights to address class imbalance
for each training epoch **do**
  for each mini-batch (x, y) **do**
    Resample audio to 16 kHz and normalize duration
    Extract input features using Wav2Vec2 feature extractor
    Compute prediction logits using M
    Compute weighted cross-entropy loss
    Update model parameters using AdamW optimizer
  end for
  Evaluate model on validation set
  Save best-performing model based on weighted F1-score
end for
Restore best checkpoint
return Fine-tuned Wav2Vec2 model

3.1.1 Audio Preprocessing and Fine-Tuning

All speech signals were processed at a sampling rate of 16 kHz so that the inputs matched the expected format of the pretrained model. In addition, the duration of each signal was normalized to a fixed maximum length of 6 s. Utterances shorter than this duration were padded, while longer utterances were truncated. This standardization step is important because transformer-based speech models require consistent batchable inputs, and it also limits unnecessary memory usage during training. Fine-tuning was performed as a supervised emotion classification task as shown in Fig. 2. In effect, the pretrained speech encoder was adapted from general Urdu speech representation learning toward the more specific objective of modeling emotional variation in speech. Through this process, the model learned to emphasize characteristics such as pitch movement, emphasis, speaking rate, and vocal intensity that are relevant to emotional expression. The model was optimized using weighted cross-entropy loss to reduce the effect of class imbalance. Let y_i denote the ground-truth class label for sample i, $\hat{p}_{i,c}$ denote the predicted probability for class c, and w_c denote the class weight for class c. The weighted cross-entropy loss is defined as:

$$\mathcal{L}_{audio} = -\sum_{i=1}^{N} w_{y_i} \log(\hat{p}_{i,y_i}) \tag{1}$$

where N represents the batch size.

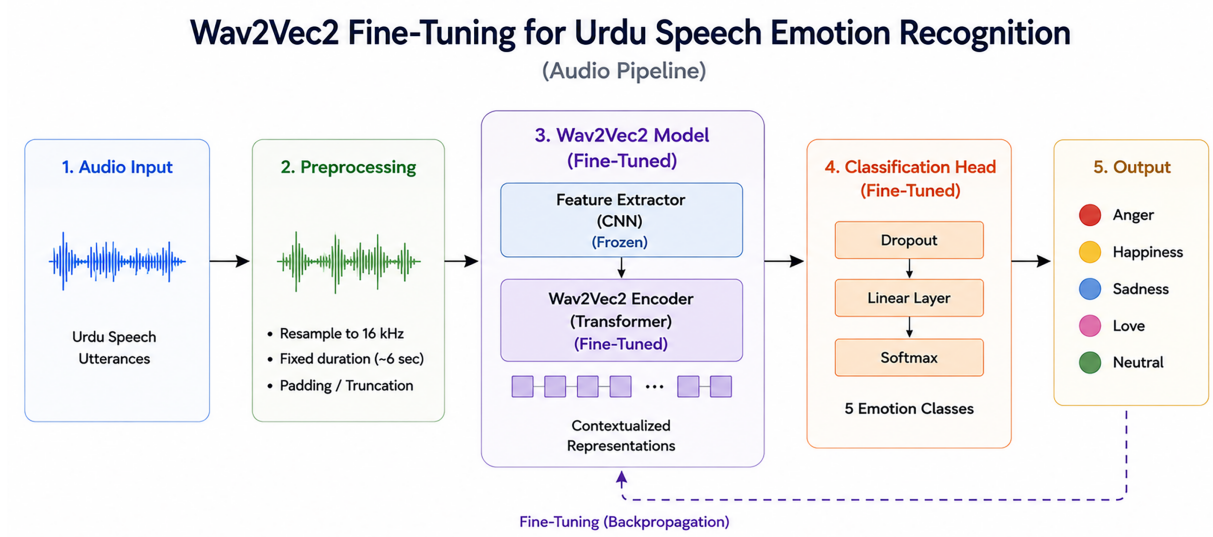

Figure 2: Audio representation learning. The Wav2Vec2 fine-tuning pipeline for Urdu speech emotion recognition. Audio inputs are preprocessed and passed through the pretrained Wav2Vec2 model, where the transformer encoder is fine-tuned for emotion classification. The extracted representations are then processed by a classification head to predict one of the five emotion classes.

3.2 Text Representation Learning

Text-based emotion recognition is challenging because emotional meaning is not always explicitly stated through words alone. The same sentence may carry different emotions depending on context, phrasing, or speaker intent. This makes contextualized language representations more suitable than shallow bag-of-words or static word embeddings.

For textual representation learning, this work used MuRIL, a transformer model developed for Indian languages and multilingual language understanding. MuRIL is well-suited for Urdu and related linguistic

contexts because it is designed to capture contextual semantics across low-resource and multilingual settings as shown in Algorithm 2. Rather than representing words in isolation, it produces context-sensitive token embeddings, which help model nuanced emotional meaning in text.

Algorithm 2: Fine-tuning MuRIL using attention-pooled multiple instance learning

Require: Training dataset $\mathcal{D}_{train}$, validation dataset $\mathcal{D}_{val}$, pretrained MuRIL model M
Ensure: Fine-tuned text emotion recognition model
Initialize pretrained MuRIL encoder with attention-pooling MIL classifier
Encode emotion labels into integer class indices
for each transcript **do**
  Split transcript into overlapping chunks using sliding-window tokenization
  Encode each chunk using MuRIL
end for
for each training epoch **do**
  for each mini-batch **do**
    Generate chunk embeddings using MuRIL encoder
    Apply attention pooling over chunk embeddings to obtain document representation
    Compute classification logits from pooled document embedding
    Compute cross-entropy loss
    Update model parameters using AdamW optimizer
  end for
  Evaluate model on validation set
  Save best-performing model based on macro F1-score
end for
Restore best checkpoint
return Fine-tuned MuRIL MIL model

3.2.1 Long-Text Handling through Multiple Instance Learning

One practical issue in transformer-based text modeling is the maximum input length. Since some transcriptions may exceed the limit of a single transformer pass, this study adopted an attention-based Multiple Instance Learning (MIL) framework. Under this design, each transcription is treated as a bag of overlapping text chunks rather than a single fixed sequence.

Each transcription was divided into chunks using a sliding window strategy with overlap. Specifically, a maximum sequence length of 256 tokens and a stride of 128 were used. Each chunk was encoded independently with MuRIL, producing a set of chunk-level embeddings. For each chunk, contextual token representations produced by MuRIL were aggregated using masked mean pooling. Let $H \in \mathbb{R}^{T \times d}$ denote the token-level hidden representations for a chunk, where T is the sequence length and d is the hidden dimension. Let m_t denote the attention mask value for token t. The chunk embedding h is computed as:

$$h = \frac{\sum_{t=1}^{T} m_t H_t}{\sum_{t=1}^{T} m_t} \tag{2}$$

These chunk-level representations were then aggregated through an attention pooling mechanism to obtain a single document-level embedding. The role of attention pooling is to assign greater importance to chunks that carry stronger emotional cues, instead of treating all chunks equally. This is particularly useful in emotion recognition, where only certain parts of an utterance may be strongly indicative of the final label.

Let $\{h_1, h_2, \ldots, h_K\}$ denote the set of chunk embeddings for a document containing K chunks. Attention scores are first computed as:

$$s_k = W_2 \tanh(W_1 h_k + b_1) + b_2 \tag{3}$$

The attention weights are then obtained using the softmax function:

$$\alpha_k = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)} \tag{4}$$

Finally, the document-level representation v is computed as the weighted sum of chunk embeddings:

$$v = \sum_{k=1}^{K} \alpha_k h_k \tag{5}$$

Thus, the text pipeline does not simply average all token or chunk representations. Instead, it learns to focus selectively on emotionally informative parts of the transcription, yielding a richer and more discriminative representation for downstream fusion. The overview of MuRIL fine-tuning is shown in Fig. 3.

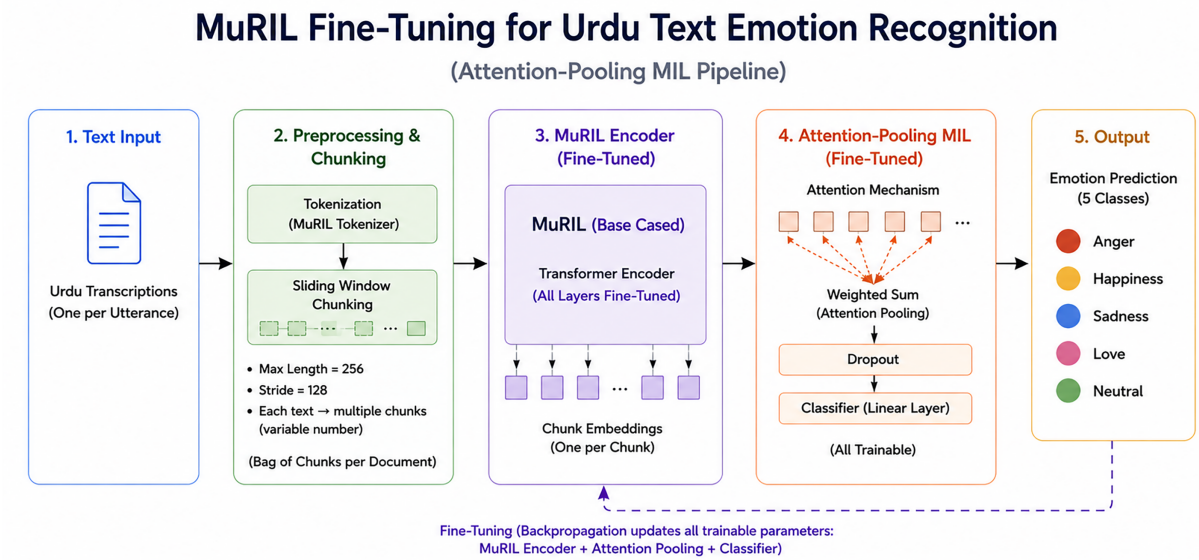

Figure 3: Text representation learning. The MuRIL fine-tuning pipeline for Urdu text emotion recognition. Urdu transcriptions are tokenized and divided into overlapping chunks, which are processed by the fine-tuned MuRIL encoder to generate chunk embeddings. These embeddings are then combined using an attention-pooling MIL mechanism before being passed to a classifier for final emotion prediction.

3.3 Embedding Extraction

After fine-tuning the unimodal models, the next stage was to convert each sample into a fixed-length dense feature vector suitable for multimodal learning. This stage serves two primary purposes. First, it separates representation learning from multimodal classification, allowing the late fusion network to operate on compact learned embeddings rather than raw audio signals or raw textual inputs. Second, it reduces the computational complexity of multimodal training by avoiding repeated transformer inference during classifier optimization. For audio embedding extraction, the fine-tuned Wav2Vec2 model was loaded and

used as a feature extractor. Instead of using the sequence classification head, the base `Wav2Vec2Model` was used so that hidden representations could be obtained directly from the encoder. Text embeddings were extracted using the fine-tuned MuRIL-based MIL model. For each transcription, chunk-level representations were first obtained using the encoder, and then attention pooling was applied to produce a single fixed-length vector per instance. The extracted audio and text embeddings are maintained as independent modality-specific representations. Rather than directly concatenating or jointly transforming these embeddings into a single shared feature space, each modality is later processed through its own dedicated classification branch before decision-level fusion is applied.

3.4 Multimodal Fusion Classification Network

The extracted audio and text embeddings originate from different pretrained models and therefore capture different forms of emotional information. Audio embeddings primarily encode acoustic and prosodic characteristics of speech, while text embeddings capture semantic and contextual information from the transcription. Instead of directly combining these heterogeneous embeddings into a shared representation space, the proposed framework processes each modality independently through separate neural classification branches. The multimodal system follows a late fusion strategy in which each modality produces its own classification logits before fusion occurs at the decision level. This design allows the model to preserve modality-specific learning characteristics while still benefiting from complementary multimodal information during final prediction.

3.4.1 Feature Normalization

Before training the multimodal classifier, the audio and text embeddings were normalized independently using z-score standardization. Separate StandardScaler transformations were fitted for the audio and text modalities. This step is important because embeddings extracted from different pretrained models may exhibit substantially different numerical distributions, magnitudes, and variances. Independent normalization helps stabilize optimization and ensures that one modality does not dominate the learning process due to scale differences alone. Since the proposed framework processes audio and text through separate modality-specific classification branches, normalization improves the numerical stability of each branch while preserving the distinct representational characteristics of both modalities. For each modality, embeddings were normalized independently using z-score standardization. Given an embedding feature value x, the normalized value $\tilde{x}$ is computed as:

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{6}$$

where μ and σ denote the mean and standard deviation computed from the training data for the corresponding modality.

3.4.2 Architecture of the Fusion Model

The proposed multimodal architecture follows a late fusion strategy in which the audio and text modalities are processed independently through separate neural classification branches before final prediction. The framework receives two inputs: speech embeddings extracted from the fine-tuned wav2vec 2.0 model and text embeddings extracted from the fine-tuned MuRIL-based Multiple Instance Learning model. Each modality is passed through its own dedicated feed-forward neural network composed of stacked multilayer perceptron (MLP) blocks. The purpose of these modality-specific branches is to adapt the pretrained embeddings to the downstream emotion recognition task while learning discriminative modality-specific

emotional representations. The basic computational unit used in both branches is an MLPBlock. Given an input feature vector x, the MLP block performs the following transformation:

$$h = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}(Wx + b)\right)\right) \tag{7}$$

where W and b denote the learnable linear transformation parameters. Each MLP block applies a linear transformation followed by Layer Normalization, GELU activation, and dropout regularization. Layer normalization improves optimization stability, GELU provides smoother nonlinear activation behavior compared to traditional activation functions such as ReLU, and dropout helps reduce overfitting during training. The audio branch receives the 1024-dimensional speech embedding and processes it through stacked MLP blocks followed by a final linear classification layer that produces logits corresponding to the target emotion classes. Similarly, the text branch processes the 768-dimensional text embedding through its own MLP-based classifier to generate an independent set of emotion logits. Rather than combining embeddings directly at the feature level, the proposed framework performs fusion at the decision level. The logits generated by the audio and text classifiers are combined using arithmetic averaging to produce the final multimodal prediction. This late fusion strategy enables the model to preserve modality-specific learning while still leveraging complementary acoustic and semantic information during emotion classification.

3.4.3 Late Fusion Strategy

The proposed multimodal framework employs a late fusion strategy for integrating speech and text information. In late fusion, each modality is processed independently and produces its own classification output before fusion occurs at the decision level. This differs from feature-level fusion approaches, where embeddings from different modalities are combined prior to classification. In the proposed architecture, speech embeddings extracted from the fine-tuned wav2vec 2.0 model are processed through an audio classification branch, while text embeddings extracted from the MuRIL-based text model are processed through a separate text classification branch. Each branch independently generates logits corresponding to the five target emotion classes.

Let z_a represent the logits generated by the audio branch and z_t represent the logits generated by the text branch. The final multimodal logits z are computed using arithmetic averaging:

$$z = \frac{z_a + z_t}{2} \tag{8}$$

The final predicted emotion class is obtained by applying the *argmax* operation to the fused logits:

$$\hat{y} = \arg\max(z) \tag{9}$$

This late fusion strategy enables both modalities to independently learn emotion-discriminative patterns while still benefiting from complementary multimodal information during final prediction. The approach is computationally simpler and more stable than complex feature-level fusion mechanisms, while still improving overall emotion recognition performance.

For comparison, two additional fusion strategies were implemented. In the early fusion approach, normalized speech and text embeddings were concatenated and provided as input to a multilayer perceptron classifier. In the gated fusion approach, a learnable gating mechanism was used to adaptively weight the modality-specific embeddings before classification. To ensure a fair comparison, all fusion strategies were trained and evaluated using the same train–test splits, optimization settings, and evaluation protocol. The only difference between the models was the fusion mechanism employed.

3.4.4 Classification

Each modality-specific branch independently produces logits corresponding to the five target emotion classes. The final multimodal prediction is obtained by averaging the logits generated by the audio and text classifiers through the late fusion mechanism described previously. The fused logits are converted into class probabilities using the softmax function:

$$P(y = c \mid z) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp(z_j)} \tag{10}$$

where C represents the total number of emotion classes.

The categorical emotion labels are encoded into integer class indices using a label encoder. Training is performed as a supervised multi-class classification problem using weighted cross-entropy loss. Class weighting is employed to address the slight imbalance in the dataset and to reduce bias toward more frequent emotion classes. The multimodal network is trained after the unimodal representation learning stage has been completed, and embeddings from both modalities have been extracted. During optimization, both modality-specific classification branches are jointly updated using the final late fusion output. Optimization is performed using the AdamW optimizer, which is well-suited for transformer-derived feature representations and deep neural network training. The overall framework, therefore, follows a staged training strategy: first learning strong unimodal representations independently, followed by multimodal classification using decision-level late fusion.

4 Experimental Setup

This section presents the dataset, hardware used, and training details used to evaluate the proposed multimodal emotion recognition framework.

4.1 Dataset

The experiments were conducted on the Urdu Multimodal Emotion Detection (UMED) dataset [12], which is specifically developed for multimodal emotion recognition in Urdu. The dataset contains 8269 unique labeled instances distributed across five emotion classes: *Anger*, *Happiness*, *Sadness*, *Love*, and *Neutral*. Each instance includes a speech utterance and its corresponding Urdu transcription. Although the original dataset contains three modalities, namely audio, text, and video, this work focuses only on the audio and text modalities. This restriction was intentional, as the primary goal of the study was to investigate whether combining speech and linguistic information alone is sufficient to improve Urdu emotion recognition.

The task is formulated as a five-class supervised classification problem. Given an utterance represented in both audio and text form, the objective is to predict its emotion label. In this setting, the audio modality contributes paralinguistic and prosodic information, while the textual modality contributes semantic and contextual information. These two sources are expected to complement one another, especially in cases where one modality alone may be ambiguous. Table 2 shows the class distribution for the dataset.

4.2 Data Preparation and Multimodal Alignment

The preparation process began with duplicate removal in the audio and text data independently. This step ensured that redundant entries did not bias training or create repeated multimodal pairs. After deduplication, the two modalities were aligned by an inner join. Only those instances that were present in both modalities were retained, which resulted in a total of 8269 unique samples, ensuring that every example used for training or testing contained both an audio signal and its corresponding transcription. The dataset

was evaluated over 10 independent experimental runs using different random seeds ranging from 43 to 52. For each run, the dataset was randomly shuffled and divided into training and test subsets using an 80/20 split. The resulting test partition was held out throughout the corresponding run and used exclusively for the final evaluation of the unimodal and multimodal models. The reported performance metrics represent the average results obtained across all 10 runs.

Table 2: Class distribution in the UMED dataset.

Emotion Class	Number of Instances
Anger (A)	2070
Happiness (H)	1774
Neutral (N)	1657
Sadness (S)	1632
Love (L)	1136

Validation data were generated independently from the training partition for each unimodal model. For Wav2Vec2 fine-tuning, 10% of the training data was reserved as a stratified validation set. For MuRIL fine-tuning, 20% of the same training data was reserved as a stratified validation set. These validation sets were used solely for model selection and training, while the held-out test set remained unseen until the final evaluation.

4.3 Hardware and Software Environment

All experiments were conducted in the Google Colaboratory environment using GPU acceleration. The training and evaluation processes were performed on an NVIDIA Tesla T4 GPU with 16 GB VRAM, supported by an Intel Xeon CPU and approximately 12–16 GB of RAM under a Linux-based Colab runtime. The implementation was developed in Python 3.12 using PyTorch 2.10.0 with CUDA 12.8 support and Torchaudio 2.10.0 for deep learning and audio processing tasks.

4.4 Training Configuration

The proposed system consists of three main components: audio representation learning, text representation learning, and multimodal fusion. For the fine-tuning process, the pre-trained wav2vec 2.0 large-XLS-R-300M model was utilized for speech emotion recognition. All audio samples were resampled to 16 kHz and normalized to a fixed duration of six seconds to ensure consistent input dimensions during training. The model was trained using the AdamW optimizer, while the low-level feature extractor remained frozen and the remaining network parameters were fine-tuned for the target emotion recognition task. The model was trained using the AdamW optimizer with a learning rate of 2×10^{-5} for 8 epochs. A per-device batch size of 2 and gradient accumulation over 8 steps were employed, resulting in an effective batch size of 16. A weight decay of 0.01 was applied, and the low-level feature extractor remained frozen during fine-tuning. In addition, early stopping with a patience of 3 epochs was employed to prevent overfitting and improve generalization performance.

The textual modality was modeled using the MuRIL transformer within an attention-based Multiple Instance Learning (MIL) framework. The transcriptions were tokenized with a maximum sequence length of 256 and a stride of 128 to efficiently handle longer text sequences through overlapping chunks. Training was performed using the AdamW optimizer with a batch size of 6 for 7 epochs, a learning rate of 2×10^{-5}, a weight decay of 0.01, and a linear learning-rate scheduler with a warmup ratio of 0.06. A dropout rate of

0.2 and gradient clipping with a maximum norm of 1.0 were employed to improve optimization stability and reduce overfitting. Finally, attention-based MIL pooling was used to aggregate chunk-level representations into a single document-level embedding for emotion classification.

The multimodal classification network was trained using embeddings extracted independently from both modalities. Audio embeddings generated by the fine-tuned wav2vec 2.0 model and text embeddings generated by the MuRIL-based text model were provided as separate inputs to the late fusion framework. Feature normalization was applied independently to both modalities prior to training in order to stabilize optimization and reduce scale-related bias between embeddings. Table 3 shows the training configuration for the fusion network.

Table 3: Multimodal fusion hyperparameters.

Parameter	Value
Input Features	Audio + Text Embeddings
Batch Size	256
Optimizer	AdamW
Loss Function	Weighted Cross-Entropy
Dropout	0.2
Normalization	StandardScaler
Fusion Strategy	Late Fusion

4.5 Evaluation Metrics

The models were evaluated using the held-out 20% test partition from each of the 10 experimental runs. The reported accuracy, precision, recall, and F1-score correspond to the average performance across all runs, providing a more robust estimate of the proposed framework's generalization performance.

5 Results and Discussion

This section presents the performance of the proposed multimodal emotion recognition framework along with unimodal baselines. The models are evaluated using accuracy, precision, recall, and F1-score. Both the fine-tuned audio and text models were evaluated on the test set along with the multimodal fusion model, and their class-wise F1-score was compared in Table 4.

Table 4: Class-wise F1-score comparison of audio, text, and multimodal models.

Model	Anger (A)	Happiness (H)	Love (L)	Neutral (N)	Sadness (S)
Audio Only	0.8047	0.6335	0.5598	0.4492	0.7886
Text Only	0.5547	0.5281	0.4359	0.4474	0.5432
Proposed Multimodal	0.8152	0.6423	0.6590	0.5482	0.7928

The audio model achieves strong performance for emotions such as *Anger* and *Sadness*, indicating that acoustic features effectively capture expressive vocal patterns. However, performance is relatively lower for *Neutral* and *Love*, suggesting that these emotions exhibit subtler acoustic variations.

The text model shows comparatively lower performance than the audio model, indicating that textual information alone may not fully capture emotional nuances, particularly for subtle or context-dependent

expressions. The multimodal model achieves the best overall performance for each class, demonstrating that combining speech and text provides complementary information for improved emotion recognition.

Fig. 4 presents the average confusion matrix of the proposed multimodal classification network over the held-out test sets from the 10 experimental runs. Each entry represents the average number of samples classified into the corresponding category across all runs, which explains the non-integer values observed in the matrix.

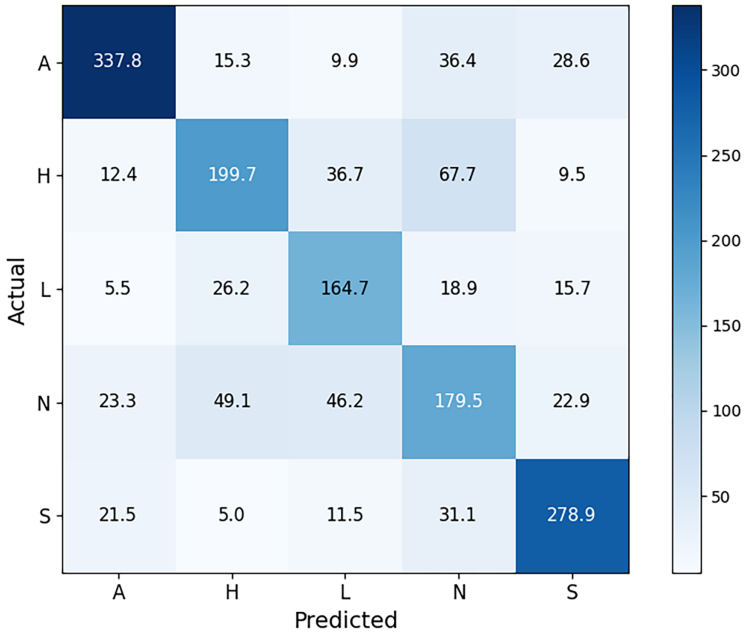

Figure 4: Average confusion matrix of the proposed multimodal classification network over 10 independent experimental runs.

To further analyze the proposed model under class imbalance, Table 5 reports the per-class precision, recall, and F1-score of the late fusion model. The values are reported as mean ± standard deviation over 10 independent runs.

Table 5: Per-class performance of the proposed late fusion model.

Emotion Class	Precision	Recall	F1-Score
Anger (A)	0.8439 ± 0.0130	0.7893 ± 0.0267	0.8153 ± 0.0126
Happiness (H)	0.6781 ± 0.0219	0.6126 ± 0.0382	0.6423 ± 0.0139
Love (L)	0.6156 ± 0.0342	0.7130 ± 0.0380	0.6590 ± 0.0132
Neutral (N)	0.5392 ± 0.0169	0.5592 ± 0.0268	0.5483 ± 0.0100
Sadness (S)	0.7854 ± 0.0243	0.8014 ± 0.0212	0.7928 ± 0.0114

The proposed late-fusion approach was compared against two other fusion strategies, namely early fusion, in which the embeddings from both modalities are concatenated before being passed to a shared classification network. The other fusion strategy evaluated was gated fusion, where modality-specific representations are adaptively combined through learnable gating mechanisms that dynamically control the contribution of each modality during multimodal integration. Experimental results showed that the

late-fusion approach achieved superior overall performance, indicating that independent modality-specific learning followed by decision-level integration was more effective. The proposed fusion strategies were evaluated over 10 independent experimental runs using different random train–test splits (random seeds 43 to 52). Table 6 reports the mean performance together with the standard deviation across all runs.

Table 6: Comparison of fusion strategies (mean ± standard deviation over 10 runs).

Fusion Strategy	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score
Early Fusion	0.6974 ± 0.0067	0.7036 ± 0.0050	0.6974 ± 0.0067	0.6989 ± 0.0056
Gated Fusion	0.6949 ± 0.0065	0.6989 ± 0.0042	0.6949 ± 0.0065	0.6958 ± 0.0055
Late Fusion	**0.7016 ± 0.0050**	**0.7078 ± 0.0036**	**0.7016 ± 0.0050**	**0.7028 ± 0.0046**

Table 6 reports weighted evaluation metrics averaged over 10 runs. In single-label multiclass classification, weighted recall is mathematically equivalent to accuracy because class-wise recalls are weighted by their class frequencies. The multimodal model outperforms both unimodal approaches, demonstrating the effectiveness of integrating complementary speech and text information through late fusion. By independently learning modality-specific emotional representations and combining predictions at the decision level, the proposed framework achieves more robust emotion recognition performance than either modality alone. All proposed models, including the unimodal baselines and the early, gated, and late fusion strategies, were trained and evaluated using the same aligned 8269-sample subset of the UMED dataset following the experimental protocol described in Section 4. The performance reported for UMEDNet was taken directly from the original publication and is included as a reference benchmark. The overall performance of the proposed methodology is compared to the unimodal pipelines, UMEDNet, along with other models in Table 7.

Table 7: Performance comparison of different models.

Model	Accuracy	Precision	Recall	Macro F1-Score
Audio Model	0.6681	0.6472	0.6532	0.6472
Text Model	0.5085	0.5027	0.5036	0.5019
XLM-R(fine-tuned)+wav2vec(fine-tuned)	0.6729	0.6684	0.6628	0.6641
mpnet(fine-tuned)+wav2vec(fine-tuned)	0.6705	0.6583	0.6604	0.6585
E5(fine-tuned)+wav2vec(fine-tuned)	0.6554	0.6445	0.6446	0.6442
UMEDNet	0.622	0.605	0.622	0.617
Proposed Multimodal Model	**0.7016**	**0.6924**	**0.6951**	**0.6915**

The proposed model achieves a significant improvement over UMEDNet and other models in all of the evaluation metrics, highlighting the effectiveness of fine-tuned transformer-based feature extraction and late multimodal fusion. The experimental results provide important insights into the effectiveness of unimodal and multimodal approaches for Urdu emotion recognition. The audio-based model achieved an accuracy of 0.6681 and a weighted F1-score of 0.6644, outperforming the text-only model. This highlights the strong role of acoustic features in capturing emotional information in speech. Class-wise analysis shows that the model performs particularly well on *Anger* (F1 = 0.8047) and *Sadness* (F1 = 0.7886). These emotions are typically associated with pronounced vocal cues such as pitch variations, intensity changes, and speech dynamics, making them easier to detect using audio signals. However, the model shows relatively lower

performance on *Neutral* (F1 = 0.4492) and *Love* (F1 = 0.5598). These emotions tend to exhibit subtle or less distinctive acoustic patterns, leading to higher confusion with other classes. In particular, neutral speech often overlaps acoustically with low-intensity emotional expressions, making it difficult to distinguish using speech features alone.

The text-based model achieved an accuracy of 0.5085 and a weighted F1-score of 0.5096, indicating significantly lower performance compared to the audio model. This behavior can be attributed to the nature of emotional expression in Urdu, where emotions are often conveyed implicitly through tone rather than explicitly through words. As a result, similar textual expressions may correspond to different emotional states depending on context.Error patterns suggest that the model struggles particularly with *Love* (F1 = 0.4359) and *Neutral* (F1 = 0.4474), which are inherently ambiguous in textual form. Additionally, overlapping vocabulary across emotion classes further contributes to misclassification, as the model relies solely on semantic cues without access to paralinguistic information.

In addition to the implicit nature of emotional expression in Urdu, several other factors may have contributed to the comparatively lower performance of the text-only model. First, Urdu's complex writing system, including spelling variations and optional diacritics, can make textual representation learning more challenging. Second, although MuRIL supports Urdu, it is a multilingual model rather than one trained exclusively on large-scale Urdu corpora, which may limit its ability to capture fine-grained linguistic and emotional nuances. Furthermore, transcriptions do not preserve prosodic cues such as intonation, stress, and speaking rate that are often essential for emotion recognition. Finally, conversational Urdu may contain code-mixed expressions or borrowed English words, introducing additional variability that can affect tokenizer coverage and representation quality. These factors collectively may explain the relatively weaker performance of the text-only model compared with the speech-based and multimodal approaches.

The proposed multimodal model achieves the best overall performance, with an accuracy of 0.7016 and an F1-score of 0.6915. This demonstrates that combining audio and text modalities leads to more robust emotion recognition than relying on a single modality alone. The improvement over unimodal models can be explained by the complementary nature of the two modalities. The audio modality captures prosodic and acoustic characteristics such as pitch variation, speaking rate, rhythm, and vocal intensity, while the textual modality contributes semantic and contextual information contained in the transcription. The proposed late fusion framework allows each modality to independently learn emotion-discriminative patterns through separate classification branches. Final predictions are then obtained through decision-level averaging of modality-specific logits. This strategy enables the model to benefit from complementary multimodal information while preserving modality-specific representational characteristics. Notably, the multimodal model improves performance on classes that are comparatively challenging for individual modalities. For example, cases where the audio signal is ambiguous can benefit from textual context, while semantically ambiguous textual expressions can be clarified through vocal cues. This complementary interaction reduces overall misclassifications and contributes to more balanced performance across emotion categories. The proposed model achieves an accuracy of 0.701, compared to 0.622 reported by UMEDNet, representing a notable improvement across all evaluation metrics. This performance gain can be attributed to several factors. First, the fine-tuning of transformer-based architectures, namely wav2vec 2.0 and MuRIL, enables more effective representation learning for both speech and text modalities. Second, the attention-based pooling mechanism used in the textual pipeline improves text representation by emphasizing emotionally informative segments within long transcriptions. Finally, the proposed late fusion framework effectively combines complementary acoustic and semantic information by integrating modality-specific classifier outputs at the decision level. The use of independent modality-specific classification branches allows the

model to preserve the representational strengths of each modality while still benefiting from multimodal integration during final prediction.

We can summarize the key observations from experimental results as follows:

- Audio features are more informative than text for emotion recognition in Urdu.
- Textual information alone is insufficient but provides complementary context.
- Multimodal fusion consistently improves performance across all metrics.
- Performance gains are observed across all evaluation metrics, indicating stable and reliable improvements.

These findings highlight the importance of multimodal approaches for low-resource languages such as Urdu, where relying on a single modality may limit the ability to capture the full spectrum of emotional expression.

Beyond improving emotion recognition performance for Urdu, the proposed framework contributes to computational modeling for low-resource multilingual systems by demonstrating how pretrained speech and language models can be effectively adapted and integrated within a unified multimodal architecture. The proposed design provides a practical foundation for deployment in affect-aware human–computer interaction systems, including intelligent virtual assistants, conversational agents, and other applications that require robust emotion understanding in low-resource language settings.

6 Conclusion and Future Directions

This study presented a multimodal deep learning framework for emotion recognition in the Urdu language by integrating speech and text modalities. The proposed approach utilized transformer-based models, namely wav2vec 2.0 for audio representation and MuRIL for text representation, combined through a late fusion strategy for multimodal classification. Experimental results demonstrated that the multimodal framework achieved superior performance compared to unimodal baselines. The audio-only model achieved an accuracy of 0.6681, while the text-only model achieved 0.5085. The proposed multimodal approach outperformed both unimodal systems, achieving an accuracy of 0.701 and an F1-score of 0.6915. Furthermore, the proposed framework surpassed the existing UMEDNet baseline, which reported an accuracy of 0.622. The results highlight the effectiveness of transformer-based representation learning together with decision-level multimodal integration for Urdu emotion recognition. The findings indicate that speech carries strong emotional cues in Urdu, while textual information provides complementary semantic context that improves overall classification performance when integrated with speech. Overall, this work demonstrates that multimodal late fusion is a promising approach for improving emotion recognition in low-resource languages such as Urdu. Future work can explore the use of more advanced or specialized embedding models for Urdu language processing. As larger and better pre-trained models become available, particularly those trained on extensive Urdu corpora, they may provide richer semantic and contextual representations. This could significantly enhance both unimodal and multimodal emotion recognition performance. The performance of deep learning models can be further improved by augmenting the dataset or collecting larger and more diverse multimodal datasets for Urdu. While the proposed late fusion framework achieved strong performance, future work may explore more advanced multimodal interaction mechanisms such as cross-modal attention or hybrid fusion strategies that combine both feature-level and decision-level information.

Acknowledgement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2601).

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2601).

Author Contributions: Conceptualization, Muhammad Sheraz, Adil Majeed, Shehzad Khalid, Yazeed Alkhrijah, Sulieman S. Alshuhri and Hasan Mujtaba; methodology, Muhammad Sheraz, Adil Majeed, Shehzad Khalid and Hasan Mujtaba; software, Muhammad Sheraz and Adil Majeed; validation, Muhammad Sheraz, Adil Majeed, Shehzad Khalid and Hasan Mujtaba; formal analysis, Muhammad Sheraz, Adil Majeed, Shehzad Khalid and Hasan Mujtaba; investigation, Muhammad Sheraz, Adil Majeed, Shehzad Khalid, Yazeed Alkhrijah, Sulieman S. Alshuhri and Hasan Mujtaba; resources, Shehzad Khalid, Yazeed Alkhrijah and Sulieman S. Alshuhri; data curation, Muhammad Sheraz and Adil Majeed; writing—original draft preparation, Muhammad Sheraz, Adil Majeed and Hasan Mujtaba; writing—review and editing, Shehzad Khalid, Yazeed Alkhrijah, Sulieman S. Alshuhri and Hasan Mujtaba; visualization, Muhammad Sheraz and Adil Majeed; supervision, Shehzad Khalid and Hasan Mujtaba; project administration, Shehzad Khalid and Hasan Mujtaba; funding acquisition, Yazeed Alkhrijah and Sulieman S. Alshuhri. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and/or analyzed during this study are available at https://zenodo.org/records/15183245.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Singh P, Srivastava R, Rana KPS, Kumar V. A multimodal hierarchical approach to speech emotion recognition from audio and text. Knowl Based Syst. 2021;229(2):107316. doi:10.1016/j.knosys.2021.107316.
2. Chutia T, Baruah N. A review on emotion detection by using deep learning techniques. Artif Intell Rev. 2024;57(8):203. doi:10.1007/s10462-024-10831-1.
3. Taj S, Mujtaba G, Daudpota SM, Mughal MH. Urdu speech emotion recognition: a systematic literature review. ACM Trans Asian Low Resour Lang Inf Process. 2023;22(7):1–33.
4. Salekin A, Eberle JW, Glenn JJ, Teachman BA, Stankovic JA. A weakly supervised learning framework for detecting social anxiety and depression. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2018;2(2):81–26. doi:10.1145/3214284.
5. Petrushin V. Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering (ANNIE); 2000 Nov 5–8; St. Louis, MI, USA.
6. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub M. A review on speech emotion recognition: case of pedagogical interaction in classroom. In: Proceedings of the International Conference on Advanced Technologies for Signal and Image Processing (ATSIP); 2017 May 22–24; Fez, Morocco. p. 1–7.
7. Szwoch M, Szwoch W. Emotion recognition for affect aware video games. In: Choraś R, editor. Image processing & communications challenges 6. Berlin/Heidelberg, Germany: Springer; 2015. doi:10.1007/978-3-319-10662-5.
8. Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T. Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun. 2008;50(6):487–503. doi:10.1016/j.specom.2008.03.012.
9. Ooi JSK, Ahmad SA, Harun HR, Chong YZ, Ali SHM. A conceptual emotion recognition framework: stress and anger analysis for car accidents. Int J Veh Saf. 2017;9(3):181–95.
10. Mozhdehi MH, Moghadam AE. Textual emotion detection utilizing a transfer learning approach. J Supercomput. 2023;79(12):13075–89. doi:10.1007/s11227-023-05168-5.
11. Shah SB, Garg S, Bourazeri A. Emotion recognition in speech by multimodal analysis of audio and text. In: Proceedings of the 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 2023 Jan 19–20; Noida, India. p. 257–63.
12. Majeed A, Mujtaba H. UMEDNet: a multimodal approach for emotion detection in the Urdu language. PeerJ Comput Sci. 2025;11(2):e2861. doi:10.7717/peerj-cs.2861.

13. Deschamps-Berger T, Lamel L, Devillers L. Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus. In: Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5.
14. Bashir MF, Javed AR, Arshad MU, Gadekallu TR, Shahzad W, Beg MO. Context-aware emotion detection from low-resource Urdu language using deep neural network. ACM Trans Asian Low Resour Lang Inf Process. 2023;22(5):1–30. doi:10.1145/3528576.
15. Adeel M, Tao ZY, Jin SY, Bejani M. Advancing low-resource Urdu speech emotion recognition: employing deep learning with unidirectional and bidirectional gated recurrent units on mel-spectrogram acoustic features. In: Proceedings of the 2025 IEEE International Conference on Pattern Recognition, Machine Vision and Artificial Intelligence (PRMVAI); 2025 Jun 20–22; Loudi, China. p. 1–9.
16. Kumar S, Haq MA, Jain A, Jason CA, Moparthi NR, Mittal N, et al. Multilayer neural network based speech emotion recognition for smart assistance. Comput Mater Contin. 2023;74(1):1524–40. doi:10.32604/cmc.2023.028631.
17. Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS. Speech emotion recognition through hybrid features and convolutional neural network. Appl Sci. 2023;13(8):4750. doi:10.3390/app13084750.
18. Singh J, Saheer LB, Faust O. Speech emotion recognition using attention model. Int J Environ Res Public Health. 2023;20(6):5140. doi:10.3390/ijerph20065140.
19. Adeel M, Tao ZY. Enhancing speech emotion recognition in Urdu using Bi-GRU networks: an in-depth analysis of acoustic features and model interpretability. In: Proceedings of the 2024 IEEE International Conference on Industrial Technology (ICIT); 2024 Mar 25–27; Bristol, UK. p. 1–6.
20. Taj S, Shaikh GM, Hassan S, Nimra. Urdu speech emotion recognition using speech spectral features and deep learning techniques. In: Proceedings of the 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET); 2023 Mar 17–18; Sukkur, Pakistan. p. 1–6.
21. Adeel M, Tao Z. Advancing speech emotion recognition for Urdu: methodological developments in low-resource contexts. In: Liu W, Wang Q, Feng J, Zhang W, editors. Proceedings of the 4th International Conference on Frontiers of Electronics, Information and Computation Technologies (ICFEICT 2024); 2024 Jun 22–24; Beijing, China.
22. Ye Z, Zuo T, Chen W, Li Y, Lu Z. Textual emotion recognition method based on ALBERT-BiLSTM model and SVM-NB classification. Soft Comput. 2023;27(8):5063–75. doi:10.1007/s00500-023-07924-4.
23. Das A, Sharif O, Hoque MM, Sarker IH. Emotion classification in a resource constrained language using transformer-based approach. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop; 2021 Jun 6–11; Online. p. 150–8. Available from: https://aclanthology.org/2021.naacl-srw.19/.
24. Kusal S, Patil S, Choudrie J, Kotecha K, Vora D. Transfer learning for emotion detection in conversational text: a hybrid deep learning approach with pre-trained embeddings. Int J Inf Technol. 2024;8(11):19. doi:10.1007/s41870-024-02027-1.
25. Majeed A, Beg MO, Arshad U, Mujtaba H. Deep-EmoRU: Mining emotions from roman Urdu text using deep learning ensemble. Multimed Tools Appl. 2022;81:43163–88. doi:10.1007/s11042-022-13147-w.
26. Ullah F, Chen X, Shah SBH, Mahfoudh S, Hassan MA, Saeed N. A novel approach for emotion detection and sentiment analysis for low resource Urdu language based on CNN-LSTM. Electronics. 2022;11(24):4096. doi:10.3390/electronics11244096.
27. Majeed A, Imtiaz U, Nseem MA, Aleem M, Shahzad W, Beg MO, et al. Extracting emotion from resource poor language through transfer learning. Multimed Tools Appl. 2025;84(19):21417–34. doi:10.1007/s11042-024-19870-w.
28. Priyasad D, Fernando T, Denman S, Sridharan S, Fookes C. Attention driven fusion for multi-modal emotion recognition. In: Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain. p. 3227–31.
29. Faria DR, Weinberg AI, Ayrosa PP. Multimodal affective communication analysis: fusing speech emotion and text sentiment using machine learning. Appl Sci. 2024;14(15):6631. doi:10.3390/app14156631.
30. Verkholyak O, Dvoynikova A, Karpov A. A bimodal approach for speech emotion recognition using audio and text. Internet Serv Inf Secur. 2021;11(1):80–96.

31. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst. 2020;33:12449–60.
32. Khanuja S, Bansal D, Mehtani S, Khosla S, Dey A, Gopalan B, et al. MuRIL: multilingual representations for indian languages. arXiv:2103.10730. 2021.