The MERSA Dataset and a Transformer-Based Approach for Speech Emotion Recognition

Enshi Zhang, Rafael Trujillo, Christian Poellabauer
Knight Foundation School of Computing & Information Sciences
Florida International University, Miami, FL 33199, USA
{ezhan004, rtruj023, cpoellab}@fiu.edu

Abstract

Research in the field of speech emotion recognition (SER) relies on the availability of comprehensive datasets to make it possible to design accurate emotion detection models. This study introduces the Multimodal Emotion Recognition and Sentiment Analysis (MERSA) dataset, which includes both natural and scripted speech recordings, transcribed text, physiological data, and self-reported emotional surveys from 150 participants collected over a two-week period. This work also presents a novel emotion recognition approach that uses a transformer-based model, integrating pre-trained wav2vec 2.0 and BERT for feature extractions and additional LSTM layers to learn hidden representations from fused representations from speech and text. Our model predicts emotions on dimensions of arousal, valence, and dominance. We trained and evaluated the model on the MSP-PODCAST dataset and achieved competitive results from the best-performing model regarding the concordance correlation coefficient (CCC). Further, this paper demonstrates the effectiveness of this model through cross-domain evaluations on both IEMOCAP and MERSA datasets.

1 Introduction

Emotions play a significant role in human interactions, as they can significantly impact our thoughts and actions (Picard, 2000). Speech-based Emotion Recognition (SER) has become increasingly popular over the last two decades due to its wide range of applications in human-computer interactions such as digital learning and mental health monitoring (Singh et al., 2023). In SER systems, emotions are commonly represented in two ways. Studies that classify emotions *categorically* often use emotions derived from Plutchik's wheel of emotions (Plutchik and Kellerman, 2013) or Ekman's six basic emotion types, which include anger, fear, disgust, happiness, sadness, and surprise, sometimes with some variations (Ekman, 1992). Emotions can also be measured along *continuous dimensions*, including valence (pleasantness), arousal (intensity), and dominance (control) (Russell and Mehrabian, 1977). The dimensional approach to studying emotions has become increasingly popular, as it can capture subtle emotional changes and represent more complex emotions than the discrete approach.

Despite the significant interest in SER in academia and industry, several challenges must be addressed. One of the significant challenges is the need for comprehensive natural datasets (Wang et al., 2022; Singh and Goel, 2022). Many datasets in this field are collected under simulated environments, which may not accurately reflect genuine emotions in real-world scenarios. Although some datasets claim to be natural, they often source their data from movies, TV shows, and online videos, raising questions about their naturalness. Another common challenge for SER is to improve the accuracy of emotion prediction. Previous research has explored emotions from various sources like text, speech, vision, and physiological data, either separately (unimodal) or together (multimodal). Combining different modalities has become a popular approach in emotion recognition, which many studies have shown to lead to more accurate prediction than the unimodal approach (Hou et al., 2022; Lee et al., 2020; Sun et al., 2021). However, there is a dearth of prior research using physiological cues for SER due to limited availability of such data.

Our work aims to address these challenges. First, we introduce the **Multimodal Emotion Recognition and Sentiment Analysis (MERSA)** dataset, which consists of data collected from 150 participants in real-world scenarios using a mobile crowdsensing approach (Ganti et al., 2011). The data was collected through daily and weekly ecological momentary assessments (EMAs). In total, we collected 37.42 hours of audio and 2,650 self-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13960–13970

reported assessments from the EMAs. Additionally, we obtained 49,550 hours of physiological data from a wrist-worn wearable device. We labeled the speech data using a novel method covering both dimensional and sentiment labels for each utterance. Second, we developed a model using a transformer-based approach by implementing wav2vec (Baevski et al., 2020) and BERT (Devlin et al., 2018) to extract acoustic and linguistic features. We trained and evaluated the model on the MSP-PODCAST (Lotfian and Busso, 2017) dataset. The best-performing model achieved competitive concordance correlation coefficient (CCC) scores on valence and arousal, and we further evaluated the model on the IEMOCAP and the MERSA datasets to demonstrate its cross-domain effectiveness and reproducibility.

Our work is divided into several sections. Section 2 covers related work, while Section 3 explains how we acquired and pre-processed the MERSA dataset. Our proposed model is described in Section 4, and Section 5 presents the datasets, models, and evaluation metrics used in our experiments, along with the results. We finally conclude our work and identify its limitations in Sections 6 and 7.

2 Related Work

This section discusses existing datasets frequently used for natural language processing (NLP) research and prior works in multimodal SER, including both categorical and dimensional approaches.

2.1 Multimodal Datasets

CH-SIMS (Yu et al., 2020) contains 2,281 video segments from real-world scenarios, with multimodal and unimodal annotations for each utterance. **CMU-MOSEI** (Zadeh et al., 2018) includes over 3,000 YouTube videos covering various topics, such as reviews and debates, and carefully selected and annotated by experts. **MSP-PODCAST** (Lotfian and Busso, 2017) is a collection of podcast recordings from audio-sharing websites featuring natural conversations on a wide range of topics and continually updated since 2017. **DECAF** (Abadi et al., 2015) captures participants' responses to music videos and movie clips, alongside physiological signals from EOG, ECG, and EMG sensors, annotated by 7 external experts. **RECOLA** (Ringeval et al., 2013) records spontaneous interactions during a remote collaborative task, including self-reports at the beginning and end of the task, involving 46 participants. **MOUD** (Pérez-Rosas et al., 2013) focuses on Spanish-language YouTube review videos, pairing each utterance with its audio, video stream, and manual transcription. **SEMAINE** (McKeown et al., 2011) provides 959 conversations from 150 participants, complete with detailed annotations and transcriptions. **YouTube** (Morency et al., 2011) encompasses a vast collection of product reviews and opinions from the social media platform, highlighting the diversity of public sentiment. Finally, the widely popular **IEMOCAP** (Busso et al., 2008) repository offers a rich dataset from ten speakers, capturing spoken communication scenarios, facial expressions, and hand movements.

2.2 Multimodal Emotion Recognition

2.2.1 Categorical

MHA (Yoon et al., 2019) utilizes an attention mechanism to selectively focus on relevant text data segments, which are then applied to corresponding audio frames to improve classification accuracy. CAN (Lee et al., 2020) integrates aligned audio and text signals by applying attention weights from one modality to the other, complementing the mutual information utilized for classification. MCSAN (Sun et al., 2021) investigates both intermodal and intra-modal interactions between audio and text. It introduces a cross-attention mechanism that allows each modality to attend to the other, refining feature representation through this guided attention. Sun (Sun et al., 2023) proposes two auxiliary tasks to improve multimodal data integration. This approach helps the network to better capture and align emotion-related features across modalities, addressing the challenge of insufficient fusion between audio and text data.

2.2.2 Dimensional

Due to the lack of labeled datasets, the work in (Li et al., 2021) utilizes unsupervised pre-training within its contrastive predictive training model and an attention-based emotion recognizer to enhance performance, demonstrating the potential of leveraging unsupervised techniques for improved accuracy. The work presented in (Triantafyllopoulos et al., 2023) introduces a novel multistage fusion method, integrating two information streams across several neural network layers. Combining data outputs from BERT and CNN, this approach tested different fusion stages to identify the most effective combination for emotion prediction. In (Atmaja

and Akagi, 2020), the authors explore a multitask learning approach, finding that text data alone is effective for valence prediction. In contrast, speech data can yield competitive results for arousal and dominance, underscoring the variable impact of different modalities on emotion recognition. The work in (Atmaja and Akagi, 2021) adopts a two-stage training process for acoustic and text features, employing an SVM for classification through a late fusion method. In (Srinivasan et al., 2022), the authors focus on improving speech representations in predicting emotions by fine-tuning wav2vec and HuBERT, with additional linguistics features from a pre-trained BERT. This approach demonstrates the effectiveness of the audio-only model using student-teacher transfer learning by achieving a high concordance correlation coefficient (CCC) on valence, with and without linguistic features. Finally, in (Wagner et al., 2023), the authors employ a transformer-based approach using HuBERT and wav2vec to achieve good results using only acoustic features. Different variants of wav2vec and HuBERT were tested to find the best-performing model.

Total enrolled participants	150
Male	95
Female	55
Another gender	0
Minimum age	19
Maximum age	51
Average age	25
Median age	24
Surveys collected	3,629
PHQ-9 surveys collected	274
Surveys with audios	2,376
Surveys with audios and emotional labels	1,181
Audio recordings collected	21,384
Total audio length (Hours)	37.42
Average audio length (Seconds)	6.3
Average words per recording	12
Electrocardiogram tests collected	2,376
Daytime physiological data length (Hours)	49,550
Sleep sessions recorded	2,751

Table 1: Statistics of the MERSA dataset.

3 The MERSA Dataset

3.1 Data Acquisition

This study recruited 150 college students and staff between October 2022 and August 2023. Participants met a member of the research team twice in our lab: once for enrollment, device pick-up, and device setup, and the second time to return the device. During the initial visit, participants received detailed information about the data collection procedures and signed a physical consent form. To ensure ethical research practices, we received approval from our school's Institutional Review Board (IRB), and every team member received training through the Collaborative Institutional Training Initiative (CITI Program). A summary of all the collected data is shown in Table 1. Each participant was required to participate for two weeks, although a few participants voluntarily extended their participation by a few days.

The participants had to complete two primary tasks during the data collection period. First, they were required to complete a daily **Emotional State Survey (ESS)** using their smartphones. The ESS consisted of 30 questions, including nine questions that required spoken responses recorded via the phone's microphone. The first audio question asked participants to describe any significant events or incidents that happened that day, and the subsequent eight audio questions required reading prescripted content aloud. The ESS also included the Positive Affect (PA) and Negative Affect (NA) Schedule (PANAS) (Watson et al., 1988), comprising 20 items equally divided between positive and negative effects. Participants responded on a Likert scale ranging from 1 ("very slightly or not at all") to 5 ("extremely"), and the questions also probed the timing of these emotions. All responses were consolidated into a single submission, with PANAS responses serving as the benchmark for data annotation. Figure 1 shows the 20 terms used in PANAS and the responses received from participants through daily ESS submissions.

Second, participants were instructed to wear the Fitbit Charge 5 wrist-worn tracker for at least 10 hours daily to collect extensive physiological data. The choice of Fitbit Charge 5 for our data collections was based on its ease of use, diverse sensor capabilities, long battery life, data quality, and data accessibility (Reid et al., 2017; Dontje et al., 2015; Diaz et al., 2015), making it superior to many other devices on the market. Further details about the device are available in its specifications and user manual (fit, 2023b,a).

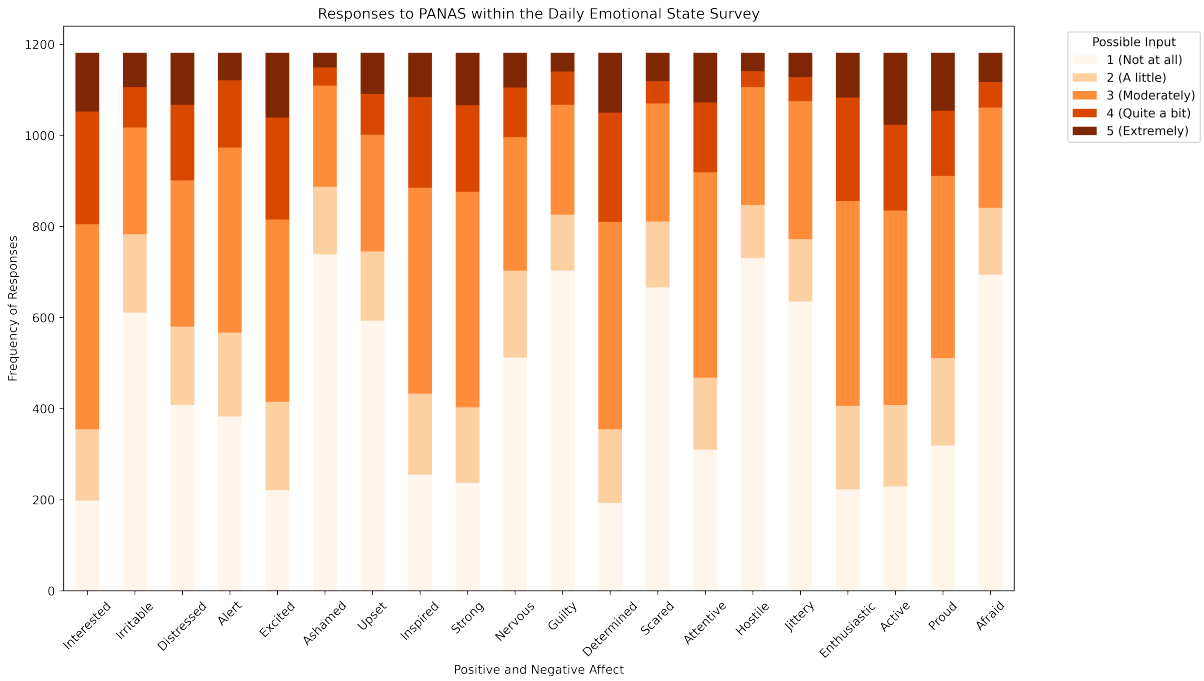

Figure 1: The 20 terms used in PANAS and the distribution of responses we received through daily ESS.

Our study also included collecting an initial demographic survey, a condensed version of the ESS excluding the PANAS, and a weekly survey incorporating the Patient Health Questionnaire-9 (PHQ-9) (Kroenke et al., 2001). To the best of our knowledge, MERSA is the most comprehensive dataset in this field regarding naturalness and comprehensiveness. It has the most significant number of real-world speakers participating (same as AVEC and SEMAINE), the most significant length of audio recordings, and an extensive amount of physiological data collected simultaneously.

To ensure the quality and quantity of data collected in a crowdsourcing study, it is essential to monitor participant compliance (Hu et al., 2020; Zhang et al., 2023). To achieve this, we created a Java-based web platform to track the study adherence of each participant, including the number of EMAs submitted and Fitbit usage. Similar to the work in (Faust et al., 2017), we measured Fitbit usage by counting the number of minutes per day that the device recorded a heartbeat, using a threshold of 10 hours (600 minutes). The visualization tool allowed us to quickly identify gaps in the sensor stream and monitor participant progress. Over the entire study period, we found the mean Fitbit usage to be 92%, with a median of 100%. We also collected 2,376 audio-based surveys from 150 students, ranging from 1 to 46, with a median of 13. This result is impressive because participants were only required to complete 14 daily surveys. The number of weekly survey submissions ranged from 1 to 10, with a median of 2, surpassing our expectations, with some participants contributing significantly more than required.

3.2 Annotation

Our annotation process distinguishes itself from methodologies like those used in the IEMOCAP and MSP-PODCAST datasets, which employ Self-Assessment Manikins (SAMs) (Lang et al., 1980; Bradley and Lang, 1994) and expert evaluations to label audio data along the dimensions of valence, arousal, and dominance. We have adopted a unique and novel method to label our data, which involves self-reported emotions from participants. This approach is supported by theories and validations from credible sources from (Mehrabian and Russell, 1974; Mehrabian, 1997; Russell and Mehrabian, 1977; Wyczesany and Ligeza, 2015). These sources have identified emotional states used in PANAS for their unique properties, which show that most of the PA and NA terms in PANAS are directly proportional to scales of valence, dominance, and arousal. Following this property, we can label the data using the following method.

To compute valence and dominance scores from PA, we directly utilize the scores for PA terms as they are, within a range of 1 to 5 for each term.

Dataset	Year	Type	Modalities	Subjects	Source	Sent	Emo	Language	Duration (hh:mm:ss)
MERSA	2023	Natural	{a, t, p}	150	Crowdsourced	✓	✓	English	37:42:00
CH-SIMS	2020	Natural	{a, t, v}	474	Online	✓	✗	Mandarin	02:19:00
CMU-MOSEI	2018	Natural	{a, t, v}	1,000	Online	✓	✓	English	65:53:36
MSP-PODCAST	2017	Natural	{a, t}	2,100	Online	✓	✓	English	237:56:00
DECAF	2015	Induced	{a, v, p}	30	Crowdsourced	✗	✓	English	44:00:00
RECOLA	2013	Natural	{a, v, p}	46	Crowdsourced	✗	✓	French	03:50:00
MOUD	2013	Natural	{a, t, v}	101	Online	✓	✗	Spanish	00:41:30
AVEC	2012	Natural	{a, v}	150	Crowdsourced	✗	✓	English	06:30:00
SEMAINE	2011	Acted	{a, v, t}	150	Crowdsourced	✗	✓	English	06:30:00
YouTube	2011	Natural	{a, t, v}	47	Online	✓	✗	English	00:23:30
IEMOCAP	2008	Acted	{a, t, v}	10	Crowdsourced	✗	✓	English	11:28:12

Table 2: Comparison of the MERSA dataset with previous sentiment analysis and emotion recognition datasets. Modality covers the subset of modalities from (a) audio, (t) text,(v) vision, (p) physiological. Duration represents total length of raw data before any preprocessing.

However, for negative affect terms, the scoring is inverted: a score of '1' (indicating minimal negativity) is reinterpreted as '5', and conversely, a score of '5' (indicating extreme negativity) is reinterpreted as '1'. This method allows the aggregated scores from PA and NA to determine the valence and dominance scores, varying from a minimum of 20 to a maximum of 100. Calculating arousal involves a more straightforward approach by directly summing the scores from PA and NA without inverting any values. This method represents the first attempt to correlate PANAS emotional states directly with the dimensional scores of valence, arousal, and dominance.

Our sentiment labeling process closely aligns with the characteristics of PANAS to label emotions as positive, negative, or neutral. We follow a similar approach to label valence by inverting the NA and summing up the scores from the PA. This results in a sentiment score range between 20 and 100. Our dataset is initially unlabeled, so we do not use the actual distribution of emotions to set our threshold. Instead, we divide the sentiment score range into three evenly distributed ranges between 20 and 100 for sentiment labeling.

3.3 Transcription

The speech data from the participants in MERSA were first transcribed using the Amazon AWS transcription service. Subsequently, to check for accuracy, these automatic transcriptions underwent manual transcription. The manual checks were thoroughly carried out to ensure transcription accuracy, resulting in a word error rate (WER) of 1.6%. An overview of the MERSA dataset and other well-known datasets used for SER and sentiment analysis is presented in Table 2.

4 Proposed Framework

To provide a clear understanding of the network architecture of our emotion recognition system, we have included Figure 2. The audio input and corresponding transcriptions were processed separately using wav2vec (Baevski et al., 2020) and BERT (Devlin et al., 2018) to extract features, and both are foundational models for speech-related applications. The wav2vec 2.0 was pre-trained on a large LibriVox dataset and fine-tuned on the whole Libri Speech dataset. It achieved the best results (WER) while using 100 times less labeled data than previous SOTA ASR systems. The model has over 95 million parameters and 12 transformer layers, with a model dimension of 768. BERT was pre-trained on a large corpus of English text and applies its self-attention mechanism to learn contextualized word embeddings. The BERT-base model has 12 transformer layers, 110 million parameters, and an embedding size 768.

Before being processed by wav2vec, the audio files were checked to ensure they were already in 16kHz and mono channel format. The raw audio is processed using wav2vec to transform it into a sequence of 768-dimensional feature vectors per time step. Given a raw audio waveform, X_{audio}, of shape $[T]$, where T is the number of time steps in the audio signal, Wav2Vec 2.0 processes X_{audio} through its layers to produce a high-dimensional

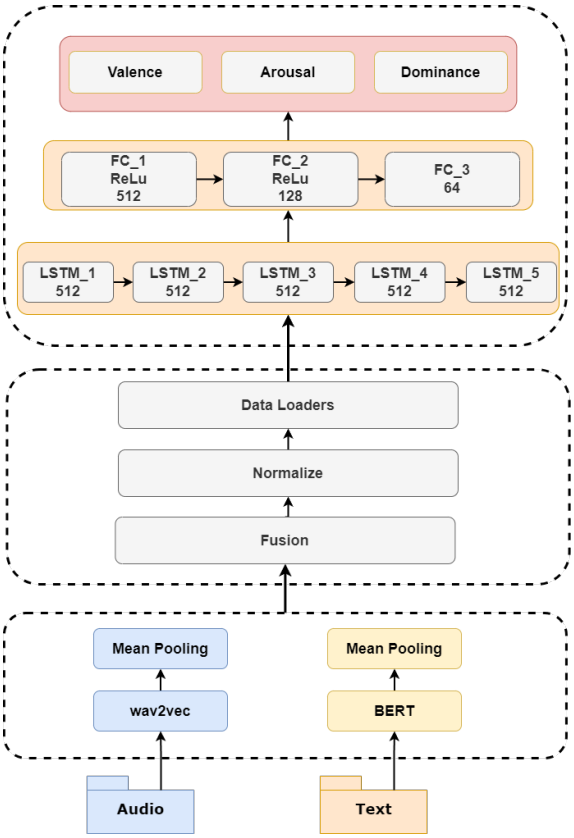

Figure 2: The overall workflow and architecture of our proposed method, consisting of 3 main modules.

representation, H_{audio}, of shape $[T, D]$, with D being the feature dimension (e.g., 768 for the base model). Mean pooling is applied across the time dimension to condense H_{audio} into a single feature vector, F_{audio}, representing the entire audio clip:

$$F_{\text{audio}} = \frac{1}{T} \sum_{t=1}^{T} H_{\text{audio}}(t)$$

This results in the final feature vector, F_{audio}, of shape $[D]$. Similarly, the BERT model extracts a sequence of 768-dimensional feature vectors from the text, and we utilize the [CLS] token representation to obtain a single 768-dimensional feature vector, resulting in F_{text} of shape $[D]$ to ensure text features are condensed to the same dimensionality.

Then, audio and textual feature vectors are concatenated to form a unified representation via early fusion before feeding them into the first dense layer. These audio and text features are fused to create a 1536-dimensional combined feature vector. For input features F_{audio} and F_{text}, both of shape $[D]$, fusion (concatenation) operation combines these two vectors along the feature dimension to produce F_{fused}:

$$F_{\text{fused}} = [F_{\text{audio}}; F_{\text{text}}]$$

resulting in F_{fused} of shape $[2D]$. We then compute mean and standard deviation statistics on these merged features and normalize them to ensure that they are on a consistent scale for model training.

Then, the model leverages a sophisticated arrangement of Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) layers followed by fully connected (dense) layers to effectively capture both the temporal dynamics and complex relationships inherent in emotional expression data. The model's core is a sequence of five LSTM layers with 512 hidden units each, configured to process the input data in a sequence-first manner. This design choice allows the network to maintain a rich contextual understanding of the input features, which is crucial for accurate emotion prediction. To mitigate overfitting, a dropout rate of 0.5 is applied within the recurrent layers. Following the recurrent layers, the model includes three dense layers. The first two layers have 256 and 128 units, respectively, and use ReLU activation functions to introduce non-linearity. They also incorporate dropout for regularization. The final dense layer is designed to produce continuous output values for each emotion dimension, aligning with the number of output emotion dimensions (e.g., 3 for valence, arousal, and dominance), and uses a linear activation function. A dropout rate of 0.5 is applied between these dense layers to improve regularization, which is consistent with the dropout strategy used in the LSTM layers.

The concordance correlation coefficient (CCC) has become the standard evaluation metric for dimensional emotion. We implement a customized loss function that optimizes for evaluation and model selection for CCC. For two random variables $\mathcal{X}$ and $\mathcal{Y}$, CCC and CCC*Loss* $\mathcal{L}$ are defined as

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

$$\mathcal{L} = 1 - CCC \quad (2)$$

The ρ in this equation is the Pearson correlation coefficient between the predicted and actual values, σ and μ are standard deviations and mean, respectively. CCC measures how well the two variables (actual values and predicted values) are aligned.

Methods	CCC		
	V	A	D
w2v2-b + BERT-b	0.516	0.490	0.494
w2v2-b + BERT-l	0.498	0.431	0.483
w2v2-l + BERT-b	0.551	0.528	0.550
w2v2-l + BERT-l	0.525	0.571	0.510
w2v2-b + BERT-b + L	0.632	0.653	0.569
w2v2-b + BERT-l + L	0.657	0.640	0.571
w2v2-l + BERT-b + L	0.620	0.545	0.633
w2v2-l + BERT-l + L	0.617	0.561	0.621

Table 3: Performance of different Transformer-based architectures trained and evaluated on the MSP-PODCAST dataset.

5 Experiments and Results

5.1 Datasets

We have trained our model to predict emotions in three dimensions (valence, arousal, and dominance) using the MSP-PODCAST (version 1.11) dataset (Lotfian and Busso, 2017). The original valence, arousal, and dominance labels range from 1 to 7, which we have normalized to a range of 0 to 1. We have used the official partition of the dataset for training, validation, and testing. Our test data is test-1, and we have 84,030 recordings for training, 19,815 recordings for validation, and 30,647 recordings for testing. To demonstrate the effectiveness of our model, we have evaluated cross-domain results on both the IEMOCAP and MERSA datasets. We have preprocessed both datasets' audio files and emotional labels to the same scale as the MSP-PODCAST. We have obtained required licenses and signed forms for using MSP-PODCAST and IEMOCAP datasets. The MERSA dataset has also been fully anonymized to remove any identity-related information.

5.2 Experimental Settings

The model was developed and executed on Google Colab Pro using an NVIDIA A100 GPU with driver version 533.104.05, CUDA version 12.2, system RAM of 83.5 GB, and GPU RAM of 40 GB. For efficient batch processing during the training phase, the model was trained with a batch size of 8 for 20 epochs. The implementation was carried out in the PyTorch framework, utilizing the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 1×10^{-4}. Performance assessment was conducted on the IEMOCAP and MERSA datasets using 5-fold and 10-fold cross-validation, respectively.

We were inspired by the research of (Wang et al., 2021) and evaluated two variants of each foundation model for feature extraction. The models we evaluated were: wav2vec 2.0-base-960 h (w2v2-b)[1], wav2vec 2.0 large (w2v2-l)[2], BERT-base (BERT-b)[3], and BERT-Large (BERT-l)[4], with and without the LSTM layer.

5.3 Results and Discussion

Table 3 compares the performance of different pre-trained models for predicting emotions in the context of the MSP-PODCAST dataset. Each model underwent training and evaluation using this specific dataset.

The combination of wav2vec 2.0 base, BERT base, and LSTM layers performed the best of all the pre-trained model configurations evaluated, achieving the highest CCC scores. Specifically, this model configuration achieved a CCC of 0.632 for valence and 0.653 for arousal, indicating its robust predictive capability for these emotional dimensions. However, a few other variants performed better than this configuration for the dominance dimension.

One consistent observation across the evaluations was that models incorporating LSTM layers performed better than those lacking this architecture across all three emotional dimensions. This outcome suggests that LSTM layers effectively improve model performance for dimensional emotion predictions. Another interesting trend was observed among models without LSTM layers. In these cases, the w2v2-l variant consistently produced more accurate predictions than the w2v2-b variant despite both models incorporating the BERT base. Notably, the w2v2-l led to a 7.75% improvement in arousal prediction accuracy, the most pronounced enhancement across the emotional dimensions assessed. These findings indicate that while LSTM layers significantly improve valence predictions, the wav2vec large variant notably boosts arousal prediction accuracy.

Table 4 outlines the results of cross-domain evaluations, highlighting the performance of baseline models across different datasets. Our model exhibited competitive performance in predicting arousal

[1] https://huggingface.co/facebook/wav2vec2-base-960h.

[2] https://huggingface.co/facebook/wav2vec2-large.

[3] https://huggingface.co/google-bert/bert-base-uncased.

[4] https://huggingface.co/google-bert/bert-large-uncased.

Dataset	Method	Modality	CCC		
			V	A	D
IEMOCAP	Two-stage SVM (Atmaja and Akagi, 2021)	a + t	0.595	0.601	0.499
	Dimensional MTL (Atmaja and Akagi, 2020)	a + t	0.446	0.594	0.486
	Multi-stage fusion (Triantafyllopoulos et al., 2023)	a + t	0.714	0.639	0.575
	RL-BERT+CNN (Srinivasan et al., 2022)	a + t	0.582	0.667	0.545
	Contrastive Unsupervised (Li et al., 2021)	a	0.752	0.752	0.691
	Pre-trained Transformer (Wagner et al., 2023)	a	0.478	0.663	0.518
	w2v2-b + BERT-b + L	a + t	0.625	0.661	0.570
MSP-PODCAST	Multi-stage fusion (Triantafyllopoulos et al., 2023)	a + t	0.714	0.639	0.575
	RL-BERT+CNN (Srinivasan et al., 2022)	a + t	0.582	0.667	0.545
	Contrastive Unsupervised (Li et al., 2021)	a	0.752	0.752	0.691
	Pre-trained Transformer (Wagner et al., 2023)	a	0.638	0.745	0.655
	w2v2-b + BERT-b + L	a + t	0.632	0.653	0.569
MERSA	**w2v2-b + BERT-b + L**	a + t	0.641	0.593	0.575

Table 4: A comparison between our top-performing model (highlighted) and other baseline models using various benchmarks.

(CCC of 0.653) and dominance (CCC of 0.569) within the MSP-PODCAST dataset, coming second only to the outcomes reported in (Li et al., 2021).

Following established experimental protocols, we subjected our model to cross-domain assessments using the IEMOCAP and MERSA datasets to validate our evaluation further. The results, shown in Table 4, demonstrate the effectiveness of the w2v-BERT-l-L model configuration. When evaluated against the IEMOCAP dataset, the model achieved CCC scores of 0.625 for valence, 0.661 for arousal, and 0.570 for dominance. This cross-domain analysis highlights the effectiveness and applicability of our model across diverse emotional datasets.

In this work, we only utilized audio and text data, excluding the physiological data from the 150 participants. Generally, our physiological data covers exercise, cardiovascular, and sleep metrics, each requiring different integration strategies for emotion recognition. For example, (Sarkar and Etemad, 2020) processed unlabeled ECG data using a self-supervised approach to learn generalized features and then used a separate network to classify emotions based on these representations with minimal labels. There are existing multimodal approaches that integrate physiological data with audio or text for emotion recognition (Katada et al., 2022; Chen et al., 2023). In future work, we plan to explore effective feature extraction methods and fusion strategies to integrate the physiological data we have collected, allowing us to investigate patterns of emotional changes in greater detail.

6 Conclusions

This paper introduces the MERSA dataset, one of the most comprehensive collections for multimodal emotion recognition and sentiment analysis. It comprises 21,384 sentences spoken by 150 real-world individuals, including both natural responses and scripted monologues. Out of these, 10,629 sentences involve self-assessments of the speakers' emotional states, which we have labeled with dimensional tags for valence, arousal, and dominance and sentiment tags for positive, negative, and neutral emotions. Additionally, the dataset contains significant physiological data that can be used for SER and other relevant studies. Although we have not utilized physiological cues in this study, we intend to use them in future studies to gain more insights into emotion recognition. The dataset and the best-performing model are currently available upon request, and we expect it to be a valuable resource for the community to pursue various NLP research initiatives. [5]

Transformers have significantly impacted various artificial intelligence tasks, including Speech Emotion Recognition (SER). Our research has validated their effectiveness by analyzing and evaluating popular transformer-based speech models for recognizing emotions in multiple dimensions. Built upon the MSP-PODCAST dataset, our model achieved competitive performance by recognizing valence with a CCC score of 0.683 and 0.590 on arousal and dominance, respectively. To further verify the effectiveness of our approach, we have evaluated its performance on the IEMOCAP and MERSA datasets, and it has also achieved competitive results.

[5] https://github.com/FIU-MOSAIC/MERSA_SER.

7 Limitations

There are limitations to both the dataset and the proposed model. The dataset's annotations rely on self-reported assessments, often considered the most accurate reflection of subjects' genuine emotions in the natural environment. However, like external expert annotations, they are not free from biases. Although we implemented a web dashboard for monitoring survey submission status, minimum wearable usage, and distribution of responses to the EMAs to improve data quality and mitigate the risk of deliberate misleading labeling, the risk cannot be entirely eliminated, potentially undermining the dataset's reliability and integrity. While most similar datasets in this field lack demographic information, MERSA includes such data, providing valuable insights. However, its demographic range is limited, as most participants are college students. This demographic homogeneity might limit the dataset's generalizability to broader populations. Nonetheless, we hope that including age information for each participant enhances the dataset's applicability for related NLP and healthcare tasks, such as depression detection.

Our model was developed and evaluated using diverse datasets that varied in their collection, annotation, and processing methodologies. While this diversity is enriching, it could introduce biases, and our preprocessing workflow may only apply to datasets with both audio and text available. Our model relies on audio signals and textual transcripts to predict dimensional labels, which may be less suitable for real-time applications requiring swift analysis and response, unlike some audio-based SER frameworks.

Acknowledgements

This research is supported by the National Science Foundation under grant 2147074. The authors would like to thank John Michael Templeton, Rahmina Rubaiat, Sajad Farrokhi, Morteza Rahimi, Maryam Songhorabadi, and Marisha Dhakal for their assistance with data collection.

The authors would like to thank David Watson, Lee Anna Clark, Jonathan Gratch, and Jeff Cohn for their insightful discussions regarding emotional labels and theories.

References

2023a. Fitbit charge 5 specifications, features and price. https://help.fitbit.com/manuals/manual_charge_5_en_US.pdf. Accessed: September 30, 2023.

2023b. Geeky wrist. https://geekywrist.com/smartwatches/fitbit-charge-5-specifications/. Accessed: September 28, 2023.

Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. 2015. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222.

Bagus Tris Atmaja and Masato Akagi. 2020. Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, 9:e17.

Bagus Tris Atmaja and Masato Akagi. 2021. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. *Speech Communication*, 126:9–21.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Sihan Chen, Jiajia Tang, Li Zhu, and Wanzeng Kong. 2023. A multi-stage dynamical fusion network for multimodal emotion recognition. *Cognitive Neurodynamics*, 17(3):671–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Keith M Diaz, David J Krupka, Melinda J Chang, James Peacock, Yao Ma, Jeff Goldsmith, Joseph E Schwartz, and Karina W Davidson. 2015. Fitbit®: An accurate and reliable device for wireless physical activity tracking. *International journal of cardiology*, 185:138–140.

Manon L Dontje, Martijn De Groot, Remko R Lengton, Cees P Van Der Schans, and Wim P Krijnen. 2015. Measuring steps with the fitbit activity tracker: an inter-device reliability study. *Journal of medical engineering & technology*, 39(5):286–290.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Louis Faust, Rachael Purta, David Hachen, Aaron Striegel, Christian Poellabauer, Omar Lizardo, and Nitesh V Chawla. 2017. Exploring compliance: Observations from a large scale fitbit study. In *Proceedings of the 2nd International Workshop on Social Sensing*, pages 55–60.

Raghu K Ganti, Fan Ye, and Hui Lei. 2011. Mobile crowdsensing: current state and future challenges. *IEEE communications Magazine*, 49(11):32–39.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mixiao Hou, Zheng Zhang, and Guangming Lu. 2022. Multi-modal emotion recognition with self-guided modality calibration. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4688–4692. IEEE.

Jiejun Hu, Kun Yang, Kezhi Wang, and Kai Zhang. 2020. A blockchain-based reward mechanism for mobile crowdsensing. *IEEE Transactions on Computational Social Systems*, 7(1):178–191.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

PJ Lang, JB Sidowski, JH Johnson, and TA Williams. 1980. Technology in mental health care delivery systems.

Yoonhyung Lee, Seunghyun Yoon, and Kyomin Jung. 2020. Multimodal Speech Emotion Recognition Using Cross Attention with Aligned Audio and Text. In *Proc. Interspeech 2020*, pages 2717–2721.

Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. 2021. Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6329–6333. IEEE.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Albert Mehrabian. 1997. Comparison of the pad and panas as models for describing emotions and for differentiating anxiety from depression. *Journal of psychopathology and behavioral assessment*, 19:331–357.

Albert Mehrabian and James A Russell. 1974. The basic emotional impact of environments. *Perceptual and motor skills*, 38(1):283–301.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Robert Plutchik and Henry Kellerman. 2013. *Theories of emotion*, volume 1. Academic press.

Ryan ER Reid, Jessica A Insogna, Tamara E Carver, Andrea M Comptour, Nicole A Bewski, Cristina Sciortino, and Ross E Andersen. 2017. Validity and reliability of fitbit activity monitors compared to actigraph gt3x+ with female adults in a free-living environment. *Journal of science and medicine in sport*, 20(6):578–582.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Pritam Sarkar and Ali Etemad. 2020. Self-supervised ecg representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554.

Jagjeet Singh, Lakshmi Babu Saheer, and Oliver Faust. 2023. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6):5140.

Youddha Beer Singh and Shivani Goel. 2022. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492:245–263.

Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. 2022. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6442–6446. IEEE.

Dekai Sun, Yancheng He, and Jiqing Han. 2023. Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Licai Sun, Bin Liu, Jianhua Tao, and Zheng Lian. 2021. Multimodal cross-and self-attention network for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4275–4279. IEEE.

Andreas Triantafyllopoulos, Uwe Reichel, Shuo Liu, Stephan Huber, Florian Eyben, and Björn W Schuller. 2023. Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Frontiers in Computer Science*, 5:1072479.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.

Miroslaw Wyczesany and Tomasz S Ligeza. 2015. Towards a constructionist approach to emotions: verification of the three-dimensional model of affect with eeg-independent component analysis. *Experimental brain research*, 233:723–733.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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 (Volume 1: Long Papers)*, pages 2236–2246.

Enshi Zhang, Rafael Trujillo, John Michael Templeton, and Christian Poellabauer. 2023. A study on mobile crowd sensing systems for healthcare scenarios. *IEEE Access*.