Multimodal Emotion Detection and Analysis from Conversational Data

Abhinay Jatoth *, Faranak Abri *, and Tien Nguyen
Department of Computer Science, San José State University, San José, USA
Email: abhinay.jatoth@sjsu.edu (A.J.); faranak.abri@sjsu.edu (F.A.);
tien.t.nguyen04@sjsu.edu (T.N.)
*Corresponding author

***Abstract*—Emotion recognition in conversations has become increasingly relevant due to its potential applications across various fields such as customer service, social media, and mental health. In this work, we explore multimodal emotion detection using both textual and audio data. Our models leverage deep learning architectures, including Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), Audio Spectrogram Transformer (AST), Wav2Vec2), Bidirectional Long Short-Term Memory (BiLSTM), and four fusion strategies that combine features from multiple modalities. We evaluate our approaches using two widely used emotion datasets, IEMOCAP and EMOV. Experimental results show that fusion models consistently outperform single-modality models, with Late Fusion achieving the highest weighted F1-Score of approximately 78% on IEMOCAP using both audio and text.**

***Keywords*—Bidirectional Encoder Representations from Transformers (BERT), conversational data, emotion recognition, fusion models, multimodal Learning, Wav2Vec2**

I. Introduction

Emotion is a powerful feeling that defines an individual's mental state and guides their actions. Researchers have utilized textual emotion detection and sentiment analysis for various applications, including customer service and chatbots [1]. In these realms, studying customer interactions and chat dialogue sentiment enables improvements in customer care experiences. Emotion detection frameworks have also been used in social media analysis and monitoring [2], which allows for the comprehension of emotions represented in social media posts and comments, as well as insights into public sentiment and opinion. Furthermore, similar frameworks have been used in psychological research and treatment [3], which provides the ability to assess emotional states and development in therapy sessions through the analysis of text or transcriptions [4]. Although recent work on emotion detection for the aforementioned applications has been promising, extracting specific emotions such as fear, nervousness, calmness, and confidence from text and speech in order to detect liars and deceivers has not been thoroughly studied.

One existing approach that has gained attention in the past few years is Emotion Recognition in Conversations (ERC), or the process of identifying emotions expressed by participants in a conversation. ERC can improve human-computer interaction, enhance customer service, and support mental health initiatives, especially in real time. However, ERC is susceptible to several challenges that make it difficult for the process to achieve a strong performance. Emotion can be expressed in different categories and is highly subjective, varying from person to person. Situational, social, and cultural factors also influence the complex and dynamic nature of emotions. Furthermore, human evaluation cannot accurately predict expressed emotions. Small utterances in a conversation, like 'Yeah!!', can signify various different possible emotions. The presence of sarcasm, emotional shift, and context can influence a conversation at any given moment. For example, the IEMOCAP dataset [5] has been annotated by six annotators. However, there have been discrepancies among different annotators in emotion category labeling, which illustrates the complexity of identifying human emotion. Accuracy is also a point of concern when it comes to ERC, especially in text and audio. These challenges not only make it difficult to build an ERC-based deep-learning model, but they also raise questions about the predicted category of emotion and its accuracy.

Identifying emotions using a dimensional model [6] that maps emotions to the nearest emotional categories is a possible technique. However, an issue arises with regard to the dataset. A dataset with proper dimensional labels is not readily available. Although IEMOCAP provides dimensional mappings ranging from 1 to 6, this is insufficient to fully explore the dataset's potential. The extraction of meaningful features plays a crucial role in emotion classification, as it directly impacts the performance of the models trained on the data. Effective features from text and audio can enhance a model's ability to distinguish between different emotional states. Once the features are extracted, various models can be employed for training. Features like Mel-Frequency Spectral Coefficients (MFCCs), chroma, pitch, and loudness have been widely used to predict emotion from audio samples. Prior studies have demonstrated that these acoustic features can effectively model perceived emotions in sound events [7]–[9]. These works show that both dimensional attributes (arousal and valence) and categorical emotion labels can be predicted from sound using machine learning approaches, with strong performances achieved by incorporating polarity and feature importance analysis. Building on this foundation, our work combines both textual and audio modalities to improve emotion recognition in conversational settings.

In this work, we explored Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN), for its ability to capture complex patterns and dependencies in

Manuscript received February 3, 2025; revised March 5, 2025; accepted May 19, 2025; published August 18, 2025.

conversational data. In addition, we utilized Large Language Models (LLMs), namely BERT (Bidirectional Encoder Representations from Transformers) [10] and RoBERTa (Robustly Optimized BERT Pretraining Approach) [11], as well as the transformer models Audio Spectogram Transformer (AST) [12] and Wav2Vec2 [13], which have demonstrated remarkable performance in various Natural Language Processing (NLP) tasks. We also employed a categorical emotion framework with labels such as happy, sad, angry, and neutral. Although dimensional models that involve valence and arousal could offer more nuanced psychological perspectives, our focus is on practical model development and benchmarking. We chose the categorical setup to remain consistent with the IEMOCAP dataset and to ensure our results are directly comparable with prior work in multimodal emotion recognition.

The remainder of this paper is structured as follows: the subsequent Section II provides an overview of existing research in emotion recognition and highlights relevant methodologies and findings. Following that, Section III discusses the datasets used in this study, including their characteristics and limitations. Section IV outlines the approaches employed in this study, detailing the feature extraction and model training processes. Section V presents the experimental setup, including the evaluation metrics, and summarizes the key findings from the experiments to provide insight into the effectiveness of the proposed methods. Finally, Section VI explains possible areas of future research in emotion recognition.

II. Related Work

Our literature review primarily focuses on different techniques for automated emotion recognition, existing challenges, and approaches adopted for ERC, using categorical models to detect emotion specifically from conversational and dialogue data, and different implementations with multimodal data. Survey papers, peer-reviewed journals, research projects, and published papers in this field have been used to address these objectives. Section II-A represents an overview of current emotion modeling and emotion datasets. Section II-B describes different methods by which emotion can be detected and dives deeper into each technique.

A. Emotion Models and Classification

1) Emotion models: Emotion is generally defined using two types of models: categorical and dimensional. The emotion labeling in the categorical model is fixed, distinct, and recognized by its respective class tags. Each category has unique characteristics and individuals can only experience one emotion at a time. On the other hand, for dimensional emotion modeling, labeling is performed on a multi-dimensional space rather than in distinct categories. Ekman *et al.* [14] proposed six basic emotion categories: anger, enjoyment, fear, sadness, disgust, and surprise, and this framework is the most widely used and well-known model for categorical emotion classification. In contrast, the dimensional model views emotions as positions in a space defined by dimensions such as valence (positive/negative), arousal (calm/excited), and dominance (controlled/submissive). Different models like Russell's circumplex [15], Thayer's energy-stress model [16], and Plutchik's "emotion wheel" [17] represent emotions within this dimensional space in different ways. Thayer's energy-stress model utilizes the two dimensions of energy and stress. Plutchik's wheel of emotions is a circular model which displays eight primary emotions and their close secondary emotions in a two-dimensional space, each pointing to its respective emotion class. Russell's circumplex model maps emotions onto a circle (circumplex shape), where opposite points represent contrasting feelings [6]. A dimensional model enables the user to choose a point in a two- or three-dimensional model, and represents a wide range of emotions in a point space.

2) Emotion datasets: Traditionally, emotion datasets focus on textual modalities, with most research centering around annotated tweets, forum posts, and movie reviews. Recently, there has been an increase in research on multimodal datasets due to their numerous use cases, especially conversational data. One of the most prominent and widely used datasets in emotion recognition research is the IEMOCAP [5] dataset. IEMOCAP (Interactive Emotional Dyadic Motion Capture), a classic multimodal dataset focusing on acted dyadic conversations, is labeled with categorical and dimensional values. Although the samples are labeled, the dimensional annotations by the evaluators are not completely accurate and there are observable discrepancies. Another dataset designed for research in emotion recognition, particularly in the context of conversations, is MELD (Multimodal EmotionLines Dataset) [18], an extension of the EmotionLines dataset featuring all textual, audio, and visual modalities from the *Friends* television series. Recently, Google has launched a large-scale dataset, GoEmotions [19], which represents more fine-grained emotions. Even though the dataset is not completely conversational, the dataset features comments extracted from Reddit conversations. These three datasets, IEMOCAP, MELD, and GoEmotions, all focus on multimodal conversational data. Other non-conversational datasets that involve basic emotions include EMOV_DB [20], EmoDB [21], RAVEDESS(Ryerson Audio-Visual Database of Emotional Expressions) [22], and CMU [23]. This literature review currently focuses primarily on conversational datasets, such as IEMOCAP, and partially on non-conversation datasets, such as EMOV, to study emotion detection.

B. Classification Approaches for Emotion Detection

1) Lexical & rule-based approach: Lexical approaches focus on analyzing a text's emotional content by examining words and phrases associated with specific emotions. On the other hand, rule-based approaches rely on predefined sets of rules and linguistic patterns to identify and classify emotions. When combined, they focus on syntactic structures, semantic relationships, and linguistic features to identify the context and classify emotion from the text. Hardik *et al.* [24] and Gaind *et al.* [25] both used a lexical and rule-based approach to identify Ekman's basic emotions in text. In [24], Hardik *et al.* identified keywords, applied rules to exclude unnecessary sentences, and considered negation words to classify the emotion class for the text. Gaind *et al.* [25] used a lexical approach with a slightly deeper analysis of textual features. They created a bag of words known as the EmotionWords Set (EWS) for Ekman's

standard emotion categories, where each word is associated with corresponding intensity levels. The authors then scored the emotions based on the degree of intensity in the text. Although approaches such as these are able to use simple NLP and syntactical rules to identify emotion labels, they struggle to classify conversational emotion, especially in utterances with fewer words.

2) Machine learning: The following section will review different implementations for emotion detection in conversation using machine learning techniques. Adopting or implementing a machine learning model involves a series of steps that include determining the dataset, selecting the model based on the target class or values to be predicted, and performing feature extraction from the dataset.

To detect emotions such as fear, nervousness, calmness, confidence, and more from text and speech using machine learning models, a comprehensive and systematic approach is followed. Firstly, publicly available annotated datasets with target emotions, such as IEMOCAP [5], are carefully selected based on their relevance, diversity, and size. This preprocessing includes removing duplicates, handling missing values, and standardizing the format of the annotations. Next, feature extraction is performed to represent emotional content in both text and speech data. For text data, a combination of lexical, syntactic, and semantic features can be considered. Lexical features involve extracting word-level statistics like term frequencies, Term Frequency-Inverse Document Frequency (TF-IDF), and n-grams to capture essential words and phrases related to emotions. Syntactic features, such as part-of-speech tags, dependency parse trees, and sentiment scores, provide information about the sentence structure and sentiment expressions. Additionally, semantic features can be derived from pre-trained word embeddings like Word2Vec, GloVe, or FastText to capture contextual meanings.

For speech data, acoustic features are extracted to capture prosodic and spectral characteristics of emotions. Common acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, and formants can be extracted from speech signals. These features are analyzed over small windows of audio to capture temporal variations in emotional expression. Once the relevant features are extracted, different machine learning models are explored, including both simple and deep learning approaches. Simple models like Support Vector Machines (SVM), Naive Bayes, and Decision Trees serve as the baseline models due to their interpretability and ease of training. Meanwhile, deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks are explored for their ability to capture complex patterns and dependencies in the data. In addition, LLMs such as BERT, GPT-3, or RoBERTa, which demonstrate better performance for NLP tasks, can also be incorporated. These models can be fine-tuned using the emotion datasets to become specialized in detecting emotions from the text. During the model training process, hyperparameter tuning and cross-validation are implemented to make the model reliable. The best-performing models are selected based on evaluation metrics like accuracy, precision, recall, F1-Score, and confusion matrix results. The most informative features can be identified using techniques like Recursive Feature Elimination (RFE) or feature importance analysis from tree-based models.

Traditional methods often struggle to capture dialogue context and conversational dynamics, where emotions can shift subtly with each turn of phrase. However, recurrence-based models, particularly LSTM, Bidirectional LSTM (BiLSTM), and Gated Recurrent Unit (GRU) can offer a deeper understanding of emotional expression in dialogue by capturing context and temporal variations within the conversation. RNNs excel at capturing the dependencies between the between words that shape emotional meaning.

Recurrence-based models, such as LSTM networks and BiLSTMs, have become widely popular for emotion detection in conversations due to their ability to capture temporal dependencies, handle sequential data, and model long-range dependencies in text. LSTM is a type of RNN designed to address the problem of vanishing gradient associated with traditional RNNs, and it excels at capturing sequential dependencies and context over extended distances. On the other hand, BiLSTM enhances the capabilities of LSTMs by processing sentences in both directions, left-to-right and right-to-left. This approach allows the model to capture both past and future context for each time step, thereby providing a more comprehensive understanding of the conversation. Refs. [26]–[28] worked on the dataset provided by Task 3 of the International Workshop on Semantic Evaluation (SemEval-2019), which consists of 15k tweet conversational pairs. Chandra *et al.* [26] chose to employ GloVe Twitter embedding with LSTM, Naive Bayes, and SVM classifiers to detect emotions in textual conversations, and reported an accuracy of 85%. On the same dataset, Rashid *et al.* [27] experimented with 3 embedding approaches, including Word2Vec, FastText, and GloVe (Global Vectors for Word Representation) with a BiLSTM network. This author achieved the highest F1 score of 69.63% with Glove Embeddings. In contrast, Ragheb *et al.* [28] applied offensive filtering, PII filtering, and language filtering on the dialogue dataset to ensure quality and relevance. The authors used a combination of lexical features, neural models, and ensembling architectures with fine-tuned BERT and LSTM. Chatterjee *et al.* [29] compared all the models used on the Task 3 dataset, and concluded that BiLSTMs/LSTMs were the most frequent models, and that GRU and CNN models were used in other implementations. Attention mechanisms, ensemble approaches, and transfer learning using BERT, ELMo (Embeddings from Language Models), and ULMFiT (Universal Language Model Fine-tuning) were popular among most implementations.

The SemEval dataset is a textual conversational pair with three to four dialogues. In order to perform ERC, a more detailed conversational dataset is necessary, such as MELD [30] or IEMOCAP [5]. Farooq *et al.* [31] proposed a novel method combining the RoBERTa model with a BiLSTM network for ERC on the MELD textual dataset. Their model achieved a weighted average F1-Score of 60.12%. Ando *et al.* [32] applied a different approach to detect emotion from speech using two datasets, namely MSP-Podcast and IEMOCAP, to implement a Listener-Adaptive (LA) model that addresses listener-dependent emotion perception. This architecture consists of an encoder, a decoder, a listener embedding layer, and adaptation layers like Adaptive Fully-Connected (AFC), Adaptive LSTM (ALSTM), and Adaptive CNN (ACNN). The author achieved a slightly higher

accuracy of 63.2% compared to [31].

All of the above implementations and research perform categorical classification for detecting emotion. Although this worked for their application, they only used a few class labels. With this method, as the number of classes increases, there will be a negative effect on accuracy. Instead of predicting the category of emotion class, Yang and Hirschberg [33], Atmaja and Akagi [34] worked on applying dimension prediction. Yang and Hirschberg [33] introduced a deep neural network model designed to track continuous changes in emotion, particularly in terms of arousal and valence. Their architecture includes CNN BiLSTM layers to handle temporal and spectral variations on the SEMAINE [35] and RECOLA [36] databases. It achieved a Concordance Correlation Coefficient (CCC) of 0.680 and 0.506 on arousal and valence, respectively. To prevent the performance of valence prediction from being affected in dimensional space, Atmaja and Akagi [34] proposed a method that integrates acoustic features with text features, thereby converting words into vectors. This approach significantly improved the prediction accuracy for valence in single-task learning. Continuous dimension prediction of arousal, valence, and dominance allows for the description of a wide range of emotion results with intensities.

GRUs are a type of RNN architecture used in deep learning. They are designed to solve the vanishing gradient problem that can occur in standard RNNs, which makes it hard for RNNs to learn and retain information over long sequences. Huddar *et al.* [37] focused on sentiment and emotion detection in conversation using multimodal data (IEMOCAP, CMU-MOSE [38]). Zadeh *et al.* [38] proposed using GRU and an attention mechanism to capture the interlocutor state and contextual information between utterances. The paper also discusses different fusion strategies, such as early fusion, model-based fusion, and late fusion. The attention-based model generally outperformed the standard baselines across different modalities and datasets. Arumugam *et al.* [39] utilized GRU to monitor a speaker's evolving mental state through their conversation, and Huang *et al.* [40] fed input features to GRU to extract global contextual information. Zhang and Xue [41] proposed a novel architecture using an autoencoder on IEMOCAP and EMODB [42] datasets to extract deep emotional features from speech. The autoencoder consists of convolution parts, instance normalization, dropout layers, and a GRU.

CNNs can be a powerful tool for emotion detection in conversations by understanding the context and capturing sequential patterns from large datasets. 1-D CNNs, Recurrent CNNs (RCNNs), and Graph CNNs (GCNNs) offer unique advantages for extracting features and predicting the target emotion. Izadkhah [43] explored the detection of multiple emotions in text data. The authors highlighted the presence of multiple emotions within conversational data and introduced a custom-built CNN architecture for this purpose. A combination of two datasets, CBET and SemEval 18, both of which contain instances with more than one emotion per sample, is used as input to the model. The combination of CNN with GloVe embedding outperformed other combinations with a Jaccard Index of 0.6463.

Nie *et al.* [44] constructed an utterance-level Graph Convolutional Network (U-GCN) with a focus on semantic correlations among utterances, and a Speaker-level GCN (S-GCN) to capture correlations between new utterances and speaker emotions. The author performed feature extraction using RoBERTa, and the S-GCN model outperformed baseline models with an F1-Score of 65.4%. Jiang *et al.* [45] performed a comparative study of CNNs and CRNNs for emotion recognition in speech. Two kinds of speech features, MFCC and GFCC, were used as input to the neural networks. CRNN uses an additional recurrent layer (GRU) to extract temporal information. After training the model, CRNN had a higher fitting degree and accuracy when compared to CNN. CRNN achieved a testing accuracy of 77.20% on MFCC features and 73.40% on GFCC features. Ghosal *et al.* [46] introduced the Dialogue Graph Convolutional Network (DialogueGCN), a graph-based approach for ERC. This method overcame the limitations of RNN-based models by using a directed graph to represent conversations, effectively capturing both sequential and speaker-level contexts. When evaluated on benchmark datasets such as IEMOCAP, AVEC, and MELD, DialogueGCN achieved an accuracy of 64.18%.

In the case of dimensional space, Li and Akagi [47] introduced a three-layer model with fuzzy inference systems for estimating these dimensions by utilizing speech features from prosodic, and spectral, and glottal waveform sources. This architecture achieved a smaller mean absolute error and a higher correlation with human evaluators. Parthasarathy and Busso [48] proposed a framework that uses Multi-Task Learning (MTL) with deep neural networks and shared hidden layers to capture the interrelation between these emotional attributes. MTL achieved significant improvements over single-task learning with a Concordance Correlation Coefficient (CCC) of 0.7635, 0.2894, and 0.7130 on arousal, valence, and dominance, respectively.

3) Other models: Along with deep Learning methods like LSTMs and CNNs, which have gained prominence in emotion detection from conversations, simple baseline machine learning models still hold value under certain circumstances, especially when the data is simple text instead of conversations. Along with Natural Language Processing (NLP), Gaind *et al.* [25] used standard classifiers such as SMO (Sequential Minimal Optimization) and J48 (a Java implementation of the C4.5 decision tree algorithm) for tweets classification, and Suhasini and Srinivasu [54] implemented classical machine learning models like Naive Bayes and K-Nearest Neighbors (KNN) on the Twitter dataset using a rule-based approach based on Russell's Circumplex model. Although simple models stand out in classifying text data, they do not perform well on conversational or multimodal data. Taking a different approach, [40] proposed an Emotion Detection Reinforcement Learning Framework (EDRLF) to detect emotions in conversations by considering both the influence of preceding Emotional States (ES) and the contextual information from previous utterances in the MELD dataset. The authors extracted textual and acoustic features separately from the dataset by utilizing GRUs, and combined them with a reinforcement learning agent (D-Q network) to perform sequential emotion detection decisions. EDRLF produced the highest w-average F1 of 60.2% on multi-modal data.

TABLE I: Model Performances for Categorical Classification Across Different Modalities: T = Text, A = Audio, V = Video

Paper	Dataset	ML Model	Modal	Metrics	Classes	Results
[26]	SemEval	GloVe+LSTM	T	Accuracy	5	85
[27]	SemEval	(Word2Vec,FastText)+BiLSTM	T	F1 Score	4	69.63
[28]	SemEval	Embeddings+BiLSTM	T	Micro F1	3	75.8
[31]	MELD	RoBERTa+BiLSTM	T	F1 Score	6	60.12
[32]	IEMOCAP	ALSTM & ACNN	A	W.Accuracy	4	61.6
[39]	IEMOCAP	VGG+GRU	A+T+V	F1-Score	5	65.4
[41]	IEMOCAP	CNN+GRU	A	UnW.Accuracy	4	71.2
[37]	IEMOCAP	(GloVe,openSmile)+GRU	A+T+V	F1-Score	5	73.3
[46]	IEMOCAP	CNN+GCN	A+V	F1-Score	6	64.18
[44]	IEMOCAP	RoBERTa+GCN	A+V	F1-Score	6	65.4
[49]	IEMOCAP	LSTM-Attn	T	F1-Score	4	63.3
[49]	IEMOCAP	AttnFusion	A+T	W.Accuracy	4	70.4
[50]	IEMOCAP	MFGCN	A+T	W.Accuracy	4	78.3
[51]	IEMOCAP	Attn+CNN	A+T	F1-Score	4	66.1
[52]	IEMOCAP	Transformer-Fusion	A+T+V	F1-Score	4	84.1
[52]	MELD	Transformer-Fusion	A+T+V	F1-Score	7	63.9
[53]	IEMOCAP	Gated-Fusion	A+T+V	W.Accuracy	4	72.39

TABLE II: Model Performances for Dimensional Prediction Across Different Modalities: T = Text, A = Audio, V = Video

Paper	Dataset	ML	Modal	Metrics	Results
Yang *et al.* [33]	SEMAINE	CNN+BiLSTM	A	CCC	(A,V)=0.68,0.50
Dim. *et al.* [34]	IEMOCAP	LSTM	A+T	Mean CCC	0.48
Li *et al.* [47]	Fujistu(1),Berlin(2)	3-Layer	A	MAE	(.16,.12)(1),(.37,.18)(2)

C. Hybrid Approach

The hybrid approach follows a combination of both lexical and machine learning approaches. Gaind *et al.* [25] combined an Emotion-Word set with SMO and J48 classifiers to achieve an accuracy of 91.7% and 85.4%, respectively. Mahima *et al.* [55] also proposed a hybrid approach for detecting multiple emotions in text and conversations from the combination of four datasets: ISEARs, MELD, EmoDB, and GoEmotions. Their hybrid model consists of manually written rules and a pre-trained model, Sentence-BERT, to identify the best emotions for each dialogue and generate multiple emotion tags. The addition of machine learning to the lexical approaches can help achieve better results when analyzing conversational data as compared to the lexical approach alone, as in Section II-B1.

A comprehensive summary of the reviewed works, for both categorical classification and dimensional prediction, is presented in Tables I and II.

III. Datasets

In this study, we used two primary datasets: IEMOCAP [5] and EMOV [20]. This work focuses on conversational emotion detection, with primary utilization of the IEMOCAP dataset. EMOV is used as an additional dataset for speech emotion recognition.

A. IEMOCAP

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset consists of approximately 12 hours of audio-visual data from 10 actors (five male and five female) who were recorded while performing scripted dialogues with a wide range of emotions, including happiness, sadness, anger, frustration, surprise, and neutral. The dataset includes both audio and visual data, as well as dialogue transcriptions. The dataset annotations include both discrete emotion labels (e.g., ”happy”, ”sad”, etc.) and dimensional emotion values like arousal, valence, and dominance. These annotations correspond to each point in the dialogue and were performed by a total of six evaluators. Each dialogue itself was labeled by three evaluators. The dataset is spread across five sessions, totaling 150 conversations and 9953 dialogues. In this work, we focus only on textual and audio data for predicting categorical labels.

There are a few challenges associated with using the IEMOCAP dataset for emotion recognition in conversations. One challenge is the large size of the dataset and the presence of noise in the audio data, which can affect the performance of the models. Additionally, the annotations provided by the evaluators may not be fully reliable, as annotations are widely mismatched when evaluated by more than one evaluator. This indicates the complexity of labeling categories for conversational data. To address these challenges, we adopt several preprocessing steps for sample selection, feature extraction, and model training.

B. EMOV

The Emotional Voices (EMOV) dataset is a collection of emotional speech recordings that can be used for training and evaluating speech models. The dataset includes recordings of four speakers (two males and two females) expressing different emotions, including neutral, sleepiness, anger, disgust, and amusement. The dataset includes 6,893 audio files with varying durations ranging from 1 to 10 s. Fig. 1 shows the distribution of emotion categories across the dataset.

IV. Methodology

In this research, we focus on textual and audio modalities for emotion detection. Separate models are trained on textual transcripts, audio, and fusions of both text and audio datasets, with key preprocessing steps and techniques for each model.

A. Dataset Preparation

The IEMOCAP dataset was first preprocessed to gather all the textual transcripts and audio dialogues from each

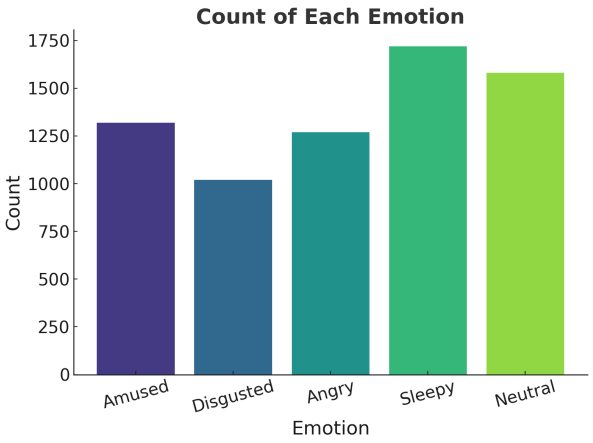

Fig. 1: Category-wise distribution of emotions in the EMOV dataset.

conversation. The final dataframe includes valuable information such as the emotion labels assigned by the three annotators, the time frame of each dialogue, audio dialogue, speaker IDs, session IDs, and conversation ID for each sample. Since each dialogue has been labeled by three different annotators, we selected the samples in which at least two annotators agreed on the annotated categorical emotion [56]. The final preprocessed dataset contains 7766 samples.

The IEMOCAP dataset provides textual transcripts for each conversation, and we aim to classify emotions using a deep learning model with this text data. We utilize multiple transformer-based models, such as BERT and RoBERTa, which possess unique features that are useful for extracting contextual embeddings.

1) Data collection and preprocessing: The dataset was preprocessed as follows:

1) **Label Selection and Mapping**: Initially, the dataset had 10 categories of emotion labels. However, labels like surprise, fear, other, and disgust were dropped due to the class size being comparatively low. To simplify classification and create a balanced dataset, some categories were merged based on similarity:
   - "Frustration" was mapped to "anger"
   - "Excited" and "happiness" were mapped to "happy"

   This label mapping strategy follows the methodology used in Yoon *et al.* [56], where similar emotional expressions were grouped to increase classification robustness and consistency.
2) **Text Cleaning**: The text data was preprocessed to remove punctuation, convert all text to lowercase, and tokenize the text using NLTK's word tokenizer.
3) **Token Filtering**: To improve the reliability of emotion classification, only sentences with 10 or more tokens were considered, and short and ambiguous sentences were left out. Many of these short samples lack clear emotional cues in both text and audio modalities. They also often consist of filler or neutral responses (e.g., "yeah," "okay," "I see") that do not contribute meaningfully to model training. Given the size of our dataset, this filtering step does not negatively impact performance and instead helps improve training stability and interpretability by increasing the signal-to-noise ratio.
4) **Label Encoding**: Labels were converted to a numerical format using `LabelEncoder` from scikit-learn, which enables compatibility with the model's output layer.

Fig. 2 shows the distribution of samples in IEMOCAP across four emotion categories after preprocessing.

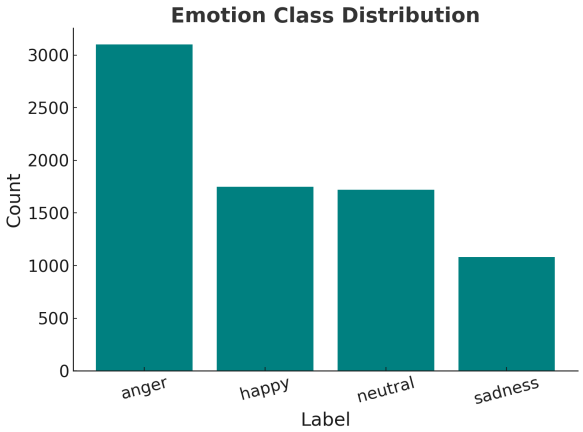

Fig. 2: Category-wise distribution of emotions in the IEMOCAP dataset after preprocessing.

2) Acoustic feature extraction: In this step, we focus on extracting a comprehensive set of acoustic features from the audio conversations from IEMOCAP and EMOV, then training the model with the extracted features. These features include:

- **Mel-Frequency Cepstral Coefficients (MFCCs)**: These coefficients represent the short-term power spectrum of sound and are widely used in speech and audio processing. Specifically, we extract 45 MFCCs and calculate their mean values.
- **Pitch**: This feature captures the perceived frequency of sound and is essential for detecting tonal variations in speech. We compute the mean and standard deviation of the pitch values.
- **Energy**: The Eoot Mean Square Energy (RMSE) measures the audio signal's amplitude and helps in understanding the intensity of the spoken words. Both the mean and standard deviation of RMSE are calculated.
- **Spectral Centroid**: This feature indicates the center of mass of the spectrum and measures the brightness of the sound. We compute both the mean and standard deviation of the spectral centroid.
- **Spectral Rolloff**: This feature represents the frequency below which a specified percentage of the total spectral energy lies. We calculate the mean spectral roll-off.
- **Spectral Bandwidth**: This feature measures the width of the spectral band and helps identify the spectrum's spread. We calculate the mean spectral bandwidth.
- **Chroma Features**: These features represent the 12 different pitch classes and provide a harmonic audio signal representation. We calculate the mean values for each of the 12 chroma features.
- **Zero-Crossing Rate (ZCR)**: ZCR tracks the rate at which the audio signal changes from positive to negative or vice versa, thereby providing insights into the

noisiness and percussiveness of the sound. We compute both the mean and standard deviation of ZCR.

We implemented our audio feature extraction pipeline using the Librosa library in Python. The audio was resampled to 16kHz and converted to mono. We extracted a total of 45 MFCCs, pitch (mean and standard deviation), Root Mean Square (RMS) energy, spectral centroid and bandwidth, spectral rolloff, chroma features (12), zero-crossing rate (mean and std), loudness (via decibel-scaled RMS), and mel-spectrogram energy. Additionally, we computed 10 Linear Predictive Coding (LPC) coefficients. All statistical features were aggregated using mean or standard deviation across the full utterance duration. This extended set of low-level descriptors provides a comprehensive representation of prosodic, spectral, and energy-related patterns relevant to emotional expression in speech.

The extracted features with and without the noise filter were then used to train with various deep learning models. Models including BiLSTM, transformer-based models, and simple neural networks are used, all implemented using the Keras library in Python. The following sections include a detailed description of each model, including its respective architectures and hyperparameters. These models are all capable of capturing temporal dependencies in the audio data, making them suitable for this task. Different models are used for audio-based, text-based, and fusion-based approaches.

B. Audio-Based Approaches

For audio-based emotion classification, three different models are used: BiLSTM, AST, and Wav2Vec2.

1) Bidirectional LSTM: Features extracted from section IV-B are used to train a BiLSTM model. The BiLSTM model consists of two didirectional LSTM layers with 128 and 64 units, respectively. LSTM is a type of RNN used to capture sequential data dependencies and reduce problems like vanishing gradients. The cells maintain information across long sequences and gates are used to control the flow, thereby maintaining a focus on context.

We added dropout layers with a rate of 0.2 after each layer in order to prevent overfitting. Our BiLSTM model also includes an Adam optimizer with a learning rate of 0.001, categorical cross-entropy as the loss function, and accuracy as the evaluation metric. Keras-Tuner is used to search for optimal configurations of LSTM units and dropout rates. Attention is used to enhance the focus on significant parts of the audio sequence.

The steps involved in training BiLSTM are as follows:

- **Data Preprocessing**: We load the audio files with a sample rate of 16 kHz and convert them to a mono channel using the Librosa library to extract relevant segments and convert them into a suitable format for feature extraction.
- **Feature Extraction**: The acoustic features are extracted using the Librosa library.
- **Model Training**: The extracted features are used to train the BiLSTM model. The model is trained to classify the emotions present in the audio conversations.
- **Evaluation**: The trained model is evaluated on a test set to determine its accuracy and F1 score. Confusion matrices are generated to visualize the model's performance across different emotions.

2) Audio Spectrogram Transformer (AST): The Audio Spectrogram Transformer (AST) [12] is a deep learning model that utilizes the transformer architecture to analyze audio data. AST applies this architecture to audio signals by converting them into spectrograms, which are visual representations of the frequency spectrum over time.

The steps involved in training AST are as follows:

- **Spectrogram Generation**: The raw audio waveform is transformed into a spectrogram.
- **Patch Embeddings**: The spectrogram is divided into smaller patches, which are then embedded into a lower-dimensional space.
- **Transformer Encoder**: These embedded patches are fed into a transformer encoder, which processes the sequence using self-attention to capture long-range dependencies.
- **Classification Head**: The output from the transformer encoder is passed through a classification head to predict emotion labels.

Fig. 3 shows the architecture of the AST model used in this experiment.

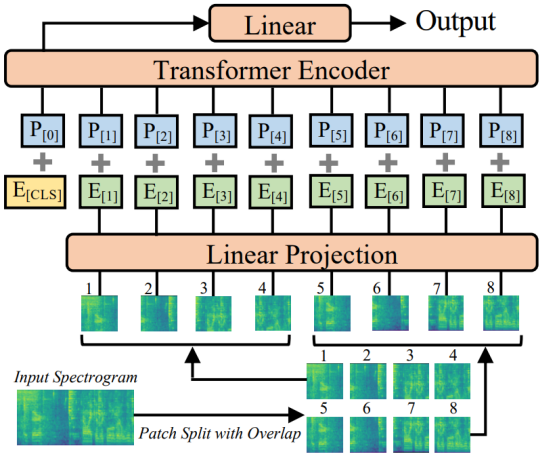

Fig. 3: Architecture of the AST model for audio classification [12].

3) Wav2Vec2: Wav2Vec2 [13] is a self-supervised learning model designed for learning speech representations from raw audio waveforms. Wav2Vec2 can be fine-tuned for specific tasks such as emotion classification. The model architecture consists of a convolutional feature encoder and a transformer context network.

The steps involved in Wav2Vec2 processing are as follows:

- **Feature Encoder**: The raw audio waveform is passed through convolutional layers to extract low-level features.
- **Context Network**: These features are processed by a transformer network to capture context and dependencies within the audio signal.
- **Quantization**: A quantization step discretizes the continuous speech representations.
- **Fine-Tuning**: The pre-trained Wav2Vec2 model is fine-tuned on labeled emotion datasets to adapt the representations for emotion classification.

Fig. 4 shows the architecture of the Wav2Vec2 model used in this experiment.

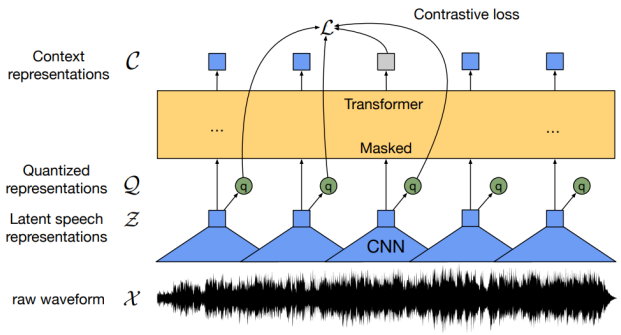

Fig. 4: Architecture of the Wav2Vec2 model for audio classification [13].

C. Text-Based Approaches

For text-based emotion classification, two different models are used: BERT and RoBERTa. These models leverage a pre-trained transformer model for feature extraction, followed by a classification layer, which is well-suited for capturing contextual information in text. Transformer-based models are selected primarily due to this ability to capture contextual information as well as semantic relationships with the help of an attention mechanism [57]. The attention mechanism not only understands individual words, but also captures their meaning by assigning different levels of importance to different words with its Query, Key, and Value mechanisms.

1) BERT (bert-base-uncased): Introduced by Google, the BERT model is well known for capturing bidirectional context in language, and can be used to achieve fine-grained emotion detection. BERT is pretrained using Next Sentence Prediction (NSP), which promotes flow between sentences, and Masked Language Modeling (MLM), which predicts the masked words in a sentence and the next sentence.

2) RoBERTa (roberta-base): RoBERTa, a robustly optimized version of BERT developed by Facebook, provides stronger contextual embeddings and effectively captures larger language patterns in emotions than BERT. RoBERTa removes NSP, which gives it the ability to focus only on MLM for better token representations.

3) Training process: Both BERT and RoBERTa are trained using two classifiers:

1) A fully connected layer is added at the end to map hidden states to the number of target classes (emotions).
2) A BiLSTM model processes the data sequence in both directions

The models are also implemented with the following techniques:

- **Learning Rate**: The models are fine-tuned on the emotion classification task with a small learning rate to avoid disrupting pretrained weights.
- **Optimizer and Loss Function**: The AdamW (Adaptive Moment Estimation with Weight Decay) optimizer is used with a learning rate of 2×10^{-5} for all models, providing stable convergence. Cross-entropy loss is used to compute the error during training.

We split the textual dataset into an 80:20 ratio for training and testing. For each model, the text samples are tokenized using the respective model tokenizers with padding and truncation to a maximum length of 128 tokens. The text is then converted into PyTorch tensors for compatibility with the models.

- **Batching**: Mini-batches with a batch size of 4 are used to balance computational efficiency and memory usage.
- **Training Loop**: Each model is trained for five epochs, with the following steps repeated in each epoch:
- **5-Fold Cross-Validation** is performed, which divides data into five subsets (folds) to train and validate the model across all subsets, improving reliability.
- **Training Process**:
  - The model predicts the labels for the training data.
  - The loss and accuracy are calculated for each batch, and gradients are computed and backpropagated.
  - Model weights are updated to minimize the training loss.

D. Fusion-Based Approaches

Fusion-based models combine multiple modalities to form a multimodal understanding of data. For IEMOCAP, audio and text features are combined at a particular stage in the fusion model pipeline. The stage at which we combine the features decides the kind of fusion. A pretrained RoBERTa model is used to extract the text embeddings, which will be used as text features, similar to what was performed in Section IV-C. Features extracted from Section IV-B are used as the audio features. Both of these features are combined in a fusion to predict the target categorical emotion.

We explore four fusion-based multimodal models: Late Fusion, Hierarchical Attention Fusion, Cross-Modal Transformer Fusion, and Gated Multimodal Fusion. Each model incorporates, encodes, and combines text and audio features in a unique way for emotion classification.

1) Late fusion: The Late Fusion model is one of the most popular forms of the fusion-based approach. In this technique, we process text and audio independently and combine their outputs later to make the final prediction. The independent processing of each modality allows each one to become specialized, and the logits are averaged to form the final classification. This is a simple and effective technique, especially when the modalities are loosely coupled, as in there is limited interaction between modalities.

The steps involved in Late Fusion are as follows:

- **Text Processing:** Contextual embeddings are extracted from textual transcripts using a pretrained RoBERTa transformer, and a dedicated text classifier is implemented as a feed-forward neural network to process the text feature.
- **Audio Processing**: Audio features are passed through an audio encoder, and a separate audio classifier processes the encoded audio features.
- **Late Fusion**: Outputs from both the text classifier and the audio classifier are combined at the logit level using a simple averaging mechanism. Then, the combined logits are used to predict the final class probabilities

Fig. 5 shows the architecture of the Late Fusion model used in this experiment.

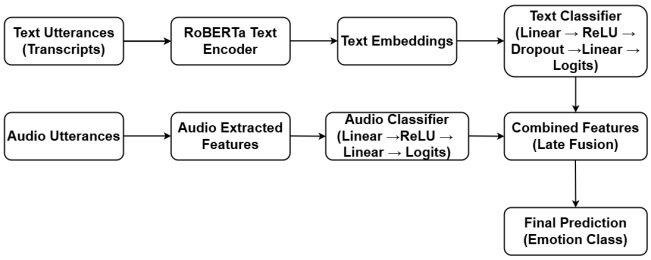

Fig. 5: Architecture of the late fusion model for audio and text classification.

2) Hierarchical attention fusion: The Hierarchical Attention Fusion model applies attention mechanisms to prioritize certain features of high importance from both modalities. Text and audio are encoded separately, and attention weights focus on important regions before concatenating the features for final classification. This method effectively captures modality-specific and cross-modal interactions and is suited for tasks where specific text phrases or audio signals dominate.

The steps involved in Hierarchical Attention Fusion are as follows:

- **Text Processing**: Contextual embeddings are extracted from textual transcripts using a pretrained RoBERTa transformer. A learnable attention mechanism assigns weights to the contextual embeddings, and attention scores are computed using a feed-forward neural network.
- **Audio Processing**: Audio features are processed in a similar way using an attention layer.
- **Fusion and Classification**: Both text and audio features are concatenated into a single multimodal representation and passed through a feed-forward classifier with ReLU activation and dropout for regularization.

Fig. 6 shows the architecture of the Hierarchical Attention Fusion model used in this experiment.

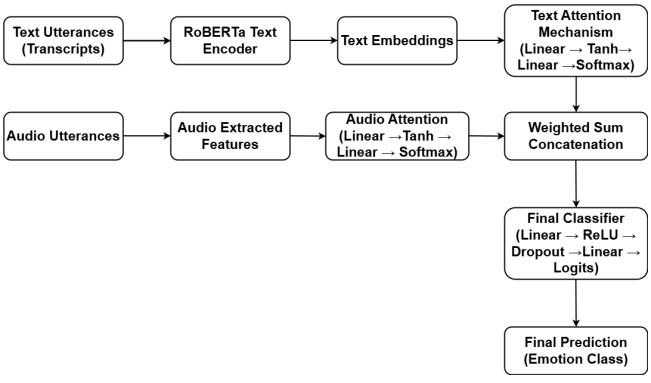

Fig. 6: Architecture of the hierarchical attention fusion model for audio and text classification.

3) Cross-modal transformer fusion: The Cross-Modal Transformer Fusion model uses a transformer encoder to fuse text and audio features. This model combines embeddings from both modalities with self-attention layers to capture complex relationships. This architecture specializes in capturing long-range dependencies and interactions [58].

The steps involved in Cross-Modal Transformer Fusion are as follows:

- **Text Processing**: Contextual embeddings are extracted from textual transcripts using a pretrained RoBERTa transformer.
- **Audio Processing**: Audio features are processed through an encoder layer.
- **Fusion and Classification**: Text and audio embeddings are concatenated and passed through the transformer encoder. A transformer encoder with eight attention heads and two layers models interactions between text and audio features. The pooled features are fused and classified into emotion categories.

Fig. 7 shows the architecture of the Cross-Modal Transformer Fusion model used in this experiment.

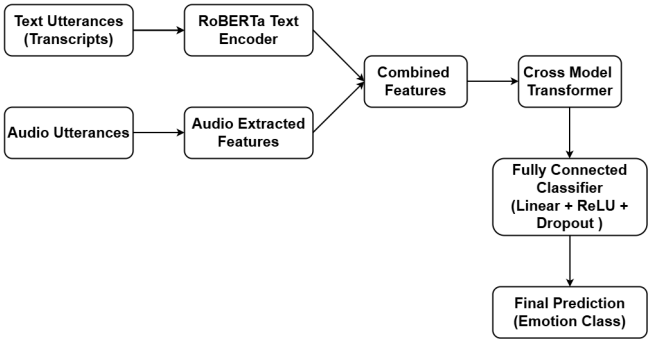

Fig. 7: Architecture of the cross-modal transformer fusionmodel for audio and text classification.

4) Gated multimodal fusion: The Gated Multimodal Fusion model employs gating mechanisms to dynamically assign weights to text and audio features. The gates decide the contribution of each modality to the final prediction. This mechanism ensures robust integration of modalities, even when one may be noisy or unreliable. This approach has been shown to improve multimodal robustness and flexibility in prior work [59].

The steps involved in Gated Multimodal Fusion are as follows:

- **Text Processing**: Contextual embeddings are extracted from textual transcripts using a pretrained RoBERTa transformer.
- **Audio Processing**: Audio features are processed through an encoder layer.
- **Gating Mechanism**: Two gating networks are used to learn the importance of both text and audio modality. Gated features are computed by element-wise multiplication of the gate values and their respective features.
- **Fusion and Classification**: The fused features are passed through a fully connected network with dropout and ReLU activation and classified into emotion categories.

Fig. 8 shows the architecture of the Gated Multimodal Fusion model used in this experiment.

V. Experiments and Results

In this section, we present the results of our experiments with various models for the task of emotion classification. Our experiments try to improve emotion classification scores compared to models mentioned in Table I.

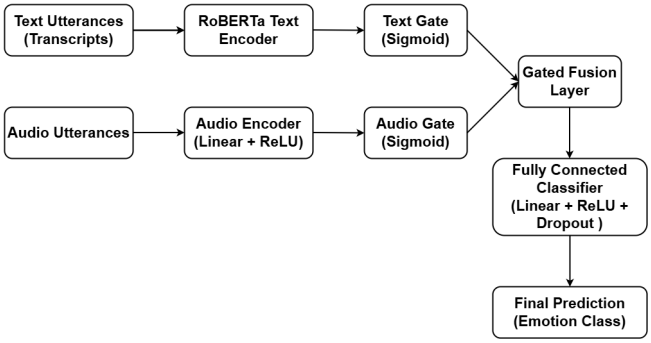

Fig. 8: Architecture of the gated multimodal fusion model for audio and text classification.

A. Evaluation Metrics

For each model, the model is evaluated using the test set after training is performed. To evaluate the performance of each model, we consider its confusion matrix and F1-Score. For our best-performing model, we report classification metrics including weighted precision, recall, and F1-Score, as well as per-class results.

1) Confusion matrix: A confusion matrix is a table used to evaluate the performance of a classification model by comparing its predicted labels with the actual labels of the samples. It is especially useful in both binary and multi-class classification tasks. Each cell in the matrix shows the number of predictions for each actual vs. predicted class combination.

For each class in a multi-class setting, the matrix allows us to define:

- **True Positive (TP)**: Correctly predicted instances of the target class.
- **True Negative (TN)**: Correctly predicted instances that do not belong to the target class.
- **False Positive (FP)**: Instances incorrectly predicted as the target class.
- **False Negative (FN)**: Instances of the target class incorrectly predicted as another class.

These values form the basis for calculating evaluation metrics such as accuracy, precision, recall, and F1-Score for each class.

2) Classification metrics: Classification metrics provide key performance indications for models' performances on predicting each emotion class in the dataset. Classification metrics include precision, recall, F1-Score, and weighted average F1-Score.

- **Precision**: Measures the accuracy of positive predictions. It is the ratio of true positives to the total predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Measures the model's ability to identify all positive instances correctly. It is the ratio of true positives to the total actual positives:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**: F1-Score is the harmonic mean of precision and recall. This balances both precision and recall, and it is useful when class distribution is imbalanced:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **The weighted average F1-Score**: Instead of calculating the F1-Score for each class and averaging them equally, the weighted average F1-Score gives more weight to classes with more samples.

$$\text{W. A. F1-Score} = \frac{\sum_{i=1}^{C} (\text{Support}_i \times \text{F1-Score}_i)}{\sum_{i=1}^{C} \text{Support}_i}$$

The performance of the models on the IEMOCAP dataset is evaluated based on four emotional categories: "happy", "sad", "neutral", and "angry". Since the IEMOCAP dataset is not balanced, we used the weighted average F1-Score as the primary metric for evaluation. The weighted average F1-Score is particularly useful in situations where the classes are imbalanced, as it ensures that each model's performance on each class is appropriately reflected in the final score.

B. Text-Based Models

For the text-based models in this study, the pretrained transformer models RoBERTa and BERT are utilized, along with a BiLSTM model. Each model is fine-tuned on the IEMOCAP dataset for the task of emotion detection using text data.

1) BERT (bert-base-uncased): BERT is known for its ability to capture bidirectional context in a language, as it performs excellently compared to many baseline models for emotion classification in text. Our BERT model was trained with the following hyperparameters and achieved a weighted F1-Score of 69.92%.

- **batch_size**: 5
- **learning_rate**: 2×10^{-5}
- **dropout**: 0.3
- **optimizer**: Adam
- **loss_function**: CrossEntropyLoss

2) RoBERTa (roberta-base): RoBERTa, a robustly optimized version of BERT, is known for its ability to capture contextual information and offer improved performance compared to BERT. Our RoBERTa model was trained with the following hyperparameters and achieved a weighted F1-Score of 72.37%.

- **batch_size**: 4
- **learning_rate**: 2×10^{-5}
- **optimizer**: AdamW
- **dropout**: 0.3
- **loss_function**: CrossEntropyLoss

3) BiLSTM: BiLSTM networks are widely used for sequential data processing tasks because they can capture long-term dependencies in both directions. Our BiLSTM model was trained with the following hyperparameters and was evaluated using 5-fold cross-validation.

- **batch_size**: 16
- **learning_rate**: 2×10^{-5}
- **optimizer**: AdamW
- **loss_function**: CrossEntropyLoss
- **dropout**: 0.2

RoBERTa with BiLSTM outperformed all the other models with a weighted F1-Score of 73.33%, and BERT with BiLSTM achieves a weighted F1-Score of 72.75%. Fig. 9 shows the confusion matrix for the BiLSTM classification of the test data with RoBERTa.

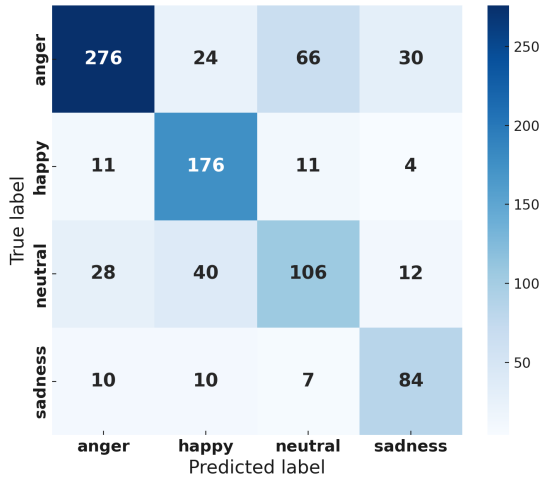

Fig. 9: Confusion matrix for BiLSTM with RoBERTa evaluated using the IEMOCAP dataset for emotion classification.

C. Audio-Based Models

For the audio-based models in this study, BiLSTM, AST, and Wav2Vec2 are trained to identify acoustic features for emotion classification. For all three models, sample utterances with a minimum token size of 10 are considered. There are a total of 4471 utterances that meet these criteria.

1) BiLSTM: All 81 acoustic features were standardized with a train and test split of 80:20. The feature vectors were then reshaped to fit the input requirements of the BiLSTM model architecture below:

- **Layers:**
  - Bi-directional LSTM: 128 units,
  - Dropout Rate: 0.2
  - Bi-directional LSTM: 64 units
  - Dropout Rate: 0.2
  - Dense: Softmax activation
  - Output Size: number of emotion categories
- **Loss Function:** Categorical cross-entropy
- **Optimizer:** Adam optimizer with default learning rate
- **Batch Size:** 32
- **Epochs:** 15

Our BiLSTM model achieved a weighted F1-Score of 62.08% on IEMOCAP and 97% on EMOV.

2) Audio Spectrogram Transformer (AST): The AST model was fine-tuned on the IEMOCAP dataset using the pre-trained model `MIT/ast-finetuned-audioset-10-10-0.4593`. The dataset was split into training and testing sets in an 80:20 ratio. We extracted audio features using the AST feature extractor and truncated audio clips or padded them to a maximum length of 10 seconds.

a) Model architecture and initialization:

- **Pretrained Model**: MIT/ast-finetuned-audioset-10-10-0.4593
- **Number of Trainable Parameters**: ∼5.3 million.
- **Output Layer**: Configured for N-class emotion classification with a `softmax` activation function.

b) Training configuration:

- **Batch Size:** 16
- **Number of Epochs:** 10
- **Learning Rate:** 5×10^{-6}
- **Weight Decay:** 0.2
- **Gradient Accumulation:** 1 step
- **Checkpointing:** Enabled for gradient optimization and best model selection

Our AST model achieved a weighted F1-Score of 64.99% on IEMOCAP, which is better than BiLSTM, and 98.96% on EMOV.

3) Wav2Vec2 Model: The Wav2Vec2 model is fine-tuned on the IEMOCAP dataset using the pre-trained `facebook/wav2vec2-base-960h` model. We split the dataset into training and testing sets with an 80:20 ratio. We extracted audio features using the Wav2Vec2 feature extractor and truncated the audio clips or padded them to a maximum length of 10 seconds.

a) Model architecture and initialization:

- **Pretrained Model**: facebook/wav2vec2-base-960h
- **Number of Trainable Parameters**: ∼95 million
- **Output Layer**: Configured for N-class emotion classification with a `softmax` activation function

b) Training configuration:

- **Batch Size:** 16
- **Number of Epochs:** 30
- **Learning Rate:** 1×10^{-5}
- **Warm-up Steps:** 50
- **Weight Decay:** 0.02
- **Gradient Accumulation:** 1 step
- **Early Stopping:** Enabled with a patience of 3 epochs
- **Checkpointing:** Enabled for gradient optimization and best model selection

Our Wav2Vec2 model achieved a weighted F1-Score of 66.81% on IEMOCAP and 99.48% on EMOV. The high performance on the EMOV dataset is likely due to its cleaner audio, less speaker variability, and smaller label set when compared to IEMOCAP. Fig. 10 shows the confusion matrix for Wav2Vec2 classification of test data.

4) Comparison: The performances of the audio-based models are shown in Table III for the EMOV dataset and the upper section of Table IV for the IEMOCAP dataset.

TABLE III: Comparison of Audio-Based Model Performances on EMOV for 5 Classes: Amused, Angry, Disgusted, Neutral, and Sleepy

ML Model	Modal	Classes	W.F1(%)
BiLSTM	Audio	5	96.8
AST	Audio	5	98.96
Wav2Vec2	Audio	5	99.48

D. Fusion-Based Models

Our fusion-based models perform the best compared to other models in the Text and Audio modals. We used 81-dimensional acoustic features, extracted as explained in Section IV-B. Four fusion-based models are implemented

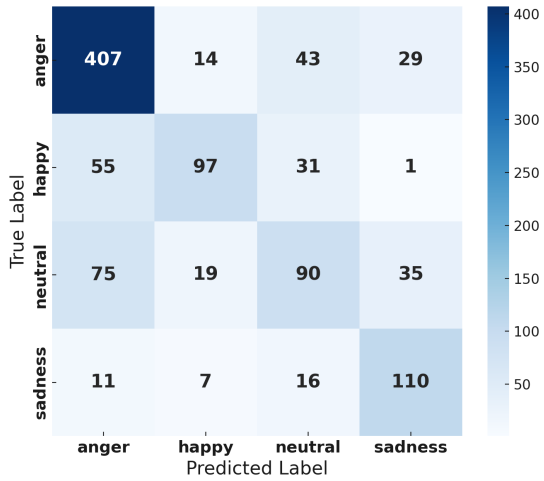

Fig. 10: Confusion matrix for Wav2Vec2 evaluated using the IEMOCAP dataset for emotion classification.

and compared: Late Fusion, Hierarchical Attention, Cross-Modal Transformer, and Gated Multimodal Fusion. Each model is trained using the following configuration:

- **Batch Size:** 16
- **Number of Epochs:** 25
- **Learning Rate:** 2×10^{-5}
- **Early Stopping:** Enabled with patience of 3 epochs.

1) Late fusion: The Late Fusion model integrates audio and text at the decision level by averaging the logits of both classifiers for each modality.

Our Late Fusion model achieved a weighted F1-Score of 77.90% on IEMOCAP. Fig. 11 shows the confusion matrix for its classification on test data.

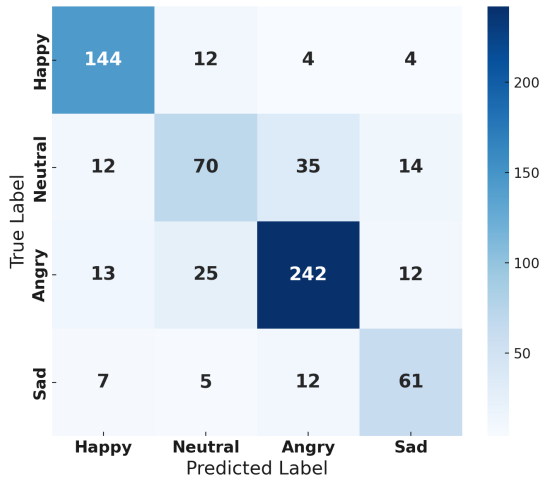

Fig. 11: Confusion matrix for late fusion evaluated using the IEMOCAP dataset for emotion classification.

2) Hierarchical attention: The Hierarchical Attention Fusion model uses attention mechanisms to process text and audio features individually, and combines them for classification.

Our Hierarchical Attention Fusion model achieved a weighted F1-Score of 75.57% on IEMOCAP.

3) Cross-modal transformer: The Cross-Modal Transformer Fusion model combines audio and text embeddings using a transformer-based architecture.

Our Cross-Modal Transformer Fusion model achieved a weighted F1-Score of 75.37% on IEMOCAP.

4) Gated multimodal: The Gated Multimodal Fusion model utilizes a gating mechanism to dynamically weight contributions of text and audio features.

Our Gated Multimodal Fusion model achieved a weighted F1-Score of 76.04% on IEMOCAP.

5) Comparison: The performance of the fusion models on the IEMOCAP dataset is presented in the lower section of Table IV. Additionally, the per-class performance of the Late Fusion model, which was identified as the best-performing fusion-bsed approach, is detailed in Table V.

6) Discussion: Fig. 12 presents a comparative analysis of training loss, validation loss, validation accuracy, and validation F1-Score across epochs for four fusion-based models: Late Fusion, Hierarchical Attention, Cross-Modal Transformer, and Gated Multimodal. These curves help illustrate how model performance evolves over time and provide insight into generalization behavior and convergence stability.

To further assess the performance of the best-performing model (Late Fusion), we computed per-class metrics including precision, recall, and F1-Score, along with weighted averages to account for class imbalance. These values are derived from the confusion matrix and provide a more detailed understanding of how the model handles each individual emotion class.

As shown in Table V, the model performs particularly well on the *happy* class, with a precision of 81.8% and a recall of 87.8%, which indicates that the Late Fusion model not only captures most of the true positives but also avoids many false positives. The *angry* class also shows strong performance with a precision of 82.6% and recall of 82.9%. In contrast, the model struggles more with the *neutral* class, where the precision drops to 62.5% and recall to 53.4%, likely due to the ambiguous and overlapping nature of neutral expressions. The *sad* class maintains a balanced performance with 67.0% precision and 71.8% recall. Overall, the model achieves a weighted precision of 77.3%, recall of 77.0%, and an F1-Score of 77.9%, reflecting strong and stable performance across all classes.

Table IV compares the weighted F1-Scores across all evaluated models. Our fusion-based approaches significantly outperform prior work on the IEMOCAP dataset. For example, the best-performing model from previous literature, Attn+CNN [51], achieves a weighted F1-Score of 66.10%. In contrast, our Late Fusion model achieves the highest score of 77.90%. Even our text-only models, such as RoBERTa+BiLSTM, perform competitively with a score of 73.33%. These results demonstrate the effectiveness of our fusion strategies and highlight the value of integrating transformer-based encoders with sequential learning layers for emotion recognition in conversations.

VI. Conclusion and Future Work

This research demonstrates the potential of fusion models trained on multimodal datasets as compared to individual modalities. Emotion recognition in conversations is a difficult task and continues to face long-standing accuracy challenges. Although text-based models are simple, fast, and perform reasonably well on the IEMOCAP dataset,

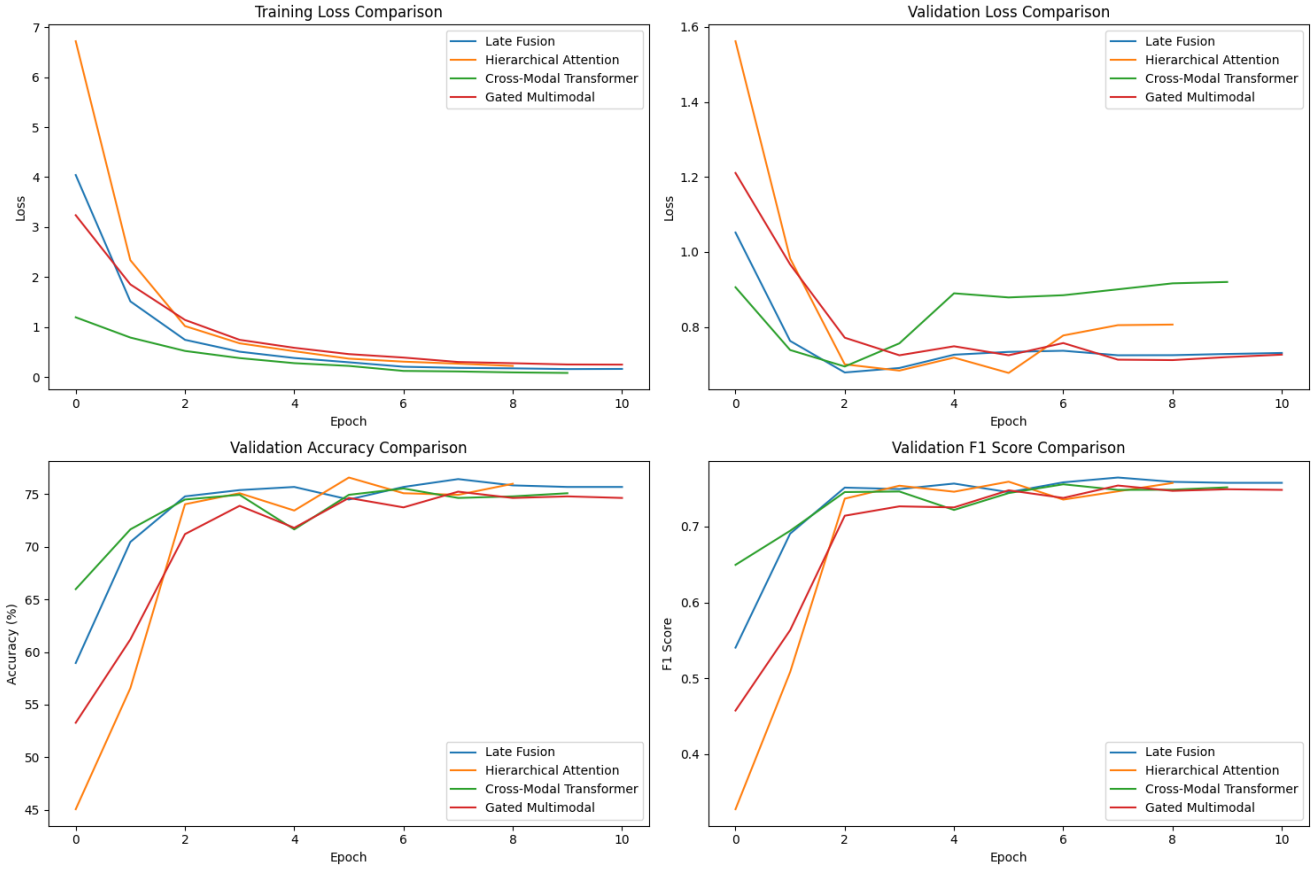

Fig. 12: Comparison of training and validation performance metrics for four fusion models evaluated using the IEMOCAP dataset for emotion classification.

TABLE IV: Comparison of Model Performance on the IEMOCAP Dataset Across Four Emotion Classes: Happy, Neutral, Angry, and Sad Modalities: A = Audio, T = Text

Model	**Modalities**	**W.F1 (%)**
RoBERTa	T	72.37
BERT	T	69.92
RoBERTa+BiLSTM	**T**	**73.33**
BERT+BiLSTM	T	72.75
BiLSTM	A	62.08
AST	A	64.99
Wav2Vec2	A	66.81
Late Fusion	**T + A**	**77.90**
Hierarchical Attention	T + A	75.57
Cross-Modal Transformer	T + A	75.37
Gated Multimodal	T + A	76.04

TABLE V: Per-Class Performance and Weighted Metrics for the Best Performing Late Fusion Model Evaluated using the IEMOCAP Dataset for Emotion Classification

Class	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Happy	81.8	87.8	84.7
Neutral	62.5	53.4	57.6
Angry	82.6	82.9	82.7
Sad	67.0	71.8	69.3
Weighted Avg.	**77.3**	**77.0**	**77.9**

fusion models consistently outperform them. For instance, models like BERT and RoBERTa with BiLSTM achieve strong results, but fusion approaches, by effectively combining information from multiple modalities, improve overall accuracy and better handle the complexity of emotion recognition. Experimental results show that while text models such as RoBERTa and BERT achieve promising weighted F1 (W.F1) scores of 73.37% and 69.92%, respectively, fusion models like Late Fusion and Gated Multimodal Fusion reach significantly higher W.F1 scores of 77.90% and 76.04%, respectively. In contrast, audio-based models underperform on IEMOCAP but perform exceptionally well on the EMOV dataset, with Wav2Vec2 reaching a W.F1 score of only 66.81% on IEMOCAP and 99.48% on EMOV. This highlights the challenges of working with complex, conversational data like IEMOCAP.

For future work, we plan to include statistical significance testing to validate performance differences between fusion models. We also aim to expand this research in several directions by incorporating additional datasets and modalities. For example, evaluating our models on larger and more diverse real-world conversational datasets will help assess generalizability and practical applicability. Cross-corpus evaluation is another important direction to better understand model robustness in varied environments. Introducing visual data, such as facial expressions, may further enhance multimodal performance. Additionally, exploring dimensional emotion models (e.g., valence-arousal representations) could offer a more nuanced understanding of emotional states compared to categorical labels. We also plan to increase the number of emotion classes, fine-tune our existing fusion models, and experiment with more advanced architectures, including large-scale pre-trained multimodal transformers. Furthermore, systematic hyperparameter tuning and ablation studies will be considered to better under-

stand their impact on model performance. Lastly, integrating domain-specific knowledge into the modeling process may further improve classification accuracy and interpretability in real-world settings.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

A.J. performed the experiments and drafted the manuscript. F.A. supervised the project, guided the methodology and analysis, validated results, revised the manuscript, and addressed reviewers' comments. T.N. contributed to experimental design refinement, result validation, manuscript revisions, and addressed reviewers' comments. All authors reviewed and approved the final version.

Funding

This research was supported by the U.S. National Science Foundation (Award#: 2319803). Opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the NSF.

Acknowledgment

We would like to thank Jade Webb from the Department of Computer Science at San José State University for her invaluable contributions to the final revisions and proofreading of this manuscript.

References

[1] S. Li, A. T. Ho, Z. Wang, and X. Zhang, "Lost in the digital wild: Hiding information in digital activities," ser. MPS '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 27–37. [Online]. Available: https://doi.org/10.1145/3267357.3267365

[2] L.-C. Chen, C.-M. Lee, and M.-Y. Chen, "Exploration of social media for sentiment analysis using deep learning," *Soft Computing*, vol. 24, pp. 8187–8197, 2020.

[3] S. B. Goldberg, N. Flemotomos, V. R. Martinez, M. J. Tanana, P. B. Kuo, B. T. Pace, J. L. Villatte, P. G. Georgiou, J. Van Epps, Z. E. Imel *et al.*, "Machine learning and natural language processing in psychotherapy research: Alliance as example use case." *Journal of counseling psychology*, vol. 67, no. 4, p. 438, 2020.

[4] B. G. Teferra, S. Borwein, D. D. DeSouza, W. Simpson, L. Rheault, and J. Rose, "Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study," *JMIR Mental Health*, vol. 9, no. 7, p. e36828, 2022.

[5] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[6] S. PS and G. Mahalakshmi, "Emotion models: a review," *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 651–657, 2017.

[7] F. Abri, L. F. Gutiérrez, A. S. Namin, D. R. Sears, and K. S. Jones, "Predicting emotions perceived from sounds," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 2057–2064.

[8] F. Abri, L. F. Gutiérrez, P. Datta, D. R. Sears, A. Siami Namin, and K. S. Jones, "A comparative analysis of modeling and predicting perceived and induced emotions in sonification," *Electronics*, vol. 10, no. 20, p. 2519, 2021.

[9] P. Krishan and F. Abri, "Classifying perceived emotions based on polarity of arousal and valence from sound events," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 2849–2856.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[14] P. Ekman, "Are there basic emotions?" *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992.

[15] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[16] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1990.

[17] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[19] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[20] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," vol. 5, 09 2005, pp. 1517–1520.

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

[23] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] S. Hardik, D. Gosai, and H. Gohil, "A review on a emotion detection and recognization from text using natural language processing," 04 2018.

[25] B. Gaind, V. Syal, and S. Padgalwar, "Emotion detection and analysis on social media," *CoRR*, vol. abs/1901.08458, 2019. [Online]. Available: http://arxiv.org/abs/1901.08458

[26] P. Chandra, M. T. Ahammed, S. Ghosh, R. H. Emon, M. Billah, M. I. Ahamad, and P. Balaji, "Contextual emotion detection in text using deep learning and big data," in *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)*, 2022, pp. 1–5.

[27] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi, and S. K. Shahzad, "Emotion detection of contextual text using deep learning," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020, pp. 1–5.

[28] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "Attention-based modeling for emotion detection and classification in textual conversations," *arXiv preprint arXiv:1906.07020*, 2019.

[29] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 39–48.

[30] S. Poria, D. Hazarika, N. Majumder, G. Naik, R. Mihalcea, and E. Cambria, "MELD: A multimodal multi-party dataset for emotion recognition in conversation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2018, pp. 527–536, https://arxiv.org/abs/1810.02508.

[31] M. Farooq, V. De Silva, H. Tibebu, and X. Shi, "Conversational emotion detection and elicitation: A preliminary study," in *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*, 2023, pp. 1–5.

[32] A. Ando, R. Masumura, H. Sato, T. Moriya, T. Ashihara, Y. Ijima, and T. Toda, "Speech emotion recognition based on listener adaptive models," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6274–6278.

[33] Z. Yang and J. Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks." in *Interspeech*, 2018, pp. 3092–3096.

[34] B. T. Atmaja and M. Akagi, "Improving valence prediction in dimensional speech emotion recognition using linguistic information," in *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2020, pp. 166–171.

[35] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroeder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2010, pp. 1079–1084.

[36] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "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)*. IEEE, 2013, pp. 1–8.

[37] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multi-modal sentiment analysis and emotion detection in conversation using rnn," 2021.

[38] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[39] B. Arumugam, S. D. Bhattacharjee, and J. Yuan, "Multimodal attentive learning for real-time explainable emotion recognition in conversations," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 1210–1214.

[40] X. Huang, M. Ren, Q. Han, X. Shi, J. Nie, W. Nie, and A.-A. Liu, "Emotion detection for conversations based on reinforcement learning framework," *IEEE MultiMedia*, vol. 28, no. 2, pp. 76–85, 2021.

[41] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE access*, vol. 9, pp. 51 231–51 241, 2021.

[42] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech." in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, vol. 5. Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.

[43] H. Izadkhah, "Detection of multiple emotions in texts using a new deep convolutional neural network," in *2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, 2022, pp. 1–6.

[44] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-gcn: incremental graph convolution network for conversation emotion detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 4471–4481, 2021.

[45] N. JIANG, J. JIA, and D. SHAO, "Comparative study of speech emotion recognition based on cnn and crnn," in *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2020, pp. 254–260.

[46] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[47] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," 2018.

[48] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Interspeech*, vol. 2017, 2017, pp. 1103–1107.

[49] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *arXiv preprint arXiv:1909.05645*, 2019.

[50] X. Qi, Y. Wen, P. Zhang, and H. Huang, "Mfgcn: Multimodal fusion graph convolutional network for speech emotion recognition," *Neurocomputing*, vol. 611, p. 128646, 2025.

[51] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, "Multimodal emotion detection via attention-based fusion of extracted facial and speech features," *Sensors*, vol. 23, no. 12, p. 5475, 2023.

[52] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[53] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," *arXiv preprint arXiv:2201.06309*, 2022.

[54] M. Suhasini and B. Srinivasu, "Emotion detection framework for twitter data using supervised classifiers," in *Data Engineering and Communication Technology*, K. S. Raju, R. Senkerik, S. P. Lanka, and V. Rajagopal, Eds. Singapore: Springer Singapore, 2020, pp. 565–576.

[55] M. A. Mahima, N. C. Patel, S. Ravichandran, N. Aishwarya, and S. Maradithaya, "A text-based hybrid approach for multiple emotion detection using contextual and semantic analysis," in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2021, pp. 1–6.

[56] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[57] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[58] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[59] J. Arevalo, T. Solorio, M. Montes-y Gomez, and F. A. Gonzalez, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.