MULTIMODAL EMOTION DETECTION IN CONVERSATIONS AND DIALOGUES: A FUSION MODEL APPROACH

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Abhinay Jatoth

Dec 2024

The Designated Project Committee Approves the Project Titled

MULTIMODAL EMOTION DETECTION IN CONVERSATIONS AND DIALOGUES: A FUSION MODEL APPROACH

by

Abhinay Jatoth

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

Dec 2024

Dr. Faranak Abri	Department of Computer Science
Dr. Navrati Saxena	Department of Computer Science
Dr. Nada Attar	Department of Computer Science

ABSTRACT

MULTIMODAL EMOTION DETECTION IN CONVERSATIONS AND DIALOGUES: A FUSION MODEL APPROACH

by Abhinay Jatoth

Emotion recognition is gaining traction due to its wide range of potential applications across different fields. With the rise of social media, chat platforms, and voice assistants, there is a vast increase in data through which humans implicitly and explicitly carry emotional cues. With new algorithms being developed for understanding the nuances of human language and emotion, businesses can tailor more personalized and empathetic service. Sentiment analysis, expresses a positive, negative, or neutral viewpoint laid the foundation of Emotion classification. Emotion classification in conversations represents the most advanced stage of classification. It is also challenging due to the existence and expression of Human emotions change from person to person in daily life. Several techniques can be used to detect emotions in text and conversations, like analyzing lexical features, machine learning, and hybrid approaches. Capturing contextual and temporal dependencies can help in the accurate prediction of Emotions in conversation. This research explores emotion classification in conversations, with textual, audio, and multimodal approaches. We experiment the effectiveness of analyzing contextual dependencies in text-based emotion detection with BERT models, also extract the acoustic features, such as MFCC, Chroma and 10 others for recognizing emotions from audio data. Additionally, implement a four different fusion techniques that combines textual and audio information to enhance the accuracy of emotion classification in conversations with various Deep Learning models.

Index Terms - ***Emotion Classification, Conversations, Acoustic Features, MultiModel, Contextual Dependencies, BERT, Fusion Models.***

ACKNOWLEDGMENTS

I am deeply indebted to my advisor, Dr. Faranak Abri. This journey wouldn't have been possible without her patient guidance through out my research. She always made time to answer my questions and provide valuable feedback that helped me for regular improvements. Her knowledge and experience were really invaluable

I would also like to express my sincere appreciation to my committee members Dr. Navrati Saxena and Dr. Nada Attar, whose exceptional teaching and mentorship have significantly shaped my academic development. Their courses provided me with a strong foundation in my field which helped me complete the project.

Finally, I would like to thank my family and friends for their constant love and support. Their understanding and encouragement have been a source of strength throughout this journey, and I am incredibly grateful for their presence in my life.

TABLE OF CONTENTS

CHAPTER

1 Introduction 1

2 Literature Review 4

2.1 Emotion Models and Classification 4

2.1.1 Emotion Models 4

2.1.2 Emotion Datasets 5

2.2 Classification Approaches for Emotion Detection 6

2.2.1 Lexical & Rule-Based Approach 6

2.2.2 Machine Learning 7

2.2.3 Hybrid Approach 13

3 Dataset 15

3.1 IEMOCAP 15

3.2 EMOV 16

4 Methodology 18

4.1 Dataset Preparation 18

4.1.1 Data Collection and Preprocessing 19

4.1.2 Acoustic Feature Extraction 19

4.2 Audio Based 21

4.2.1 Bi-directional LSTM Model : 21

4.2.2 Audio Spectrogram Transformer (AST) 22

4.2.3 Wav2Vec2 23

4.3 Text Based 24
4.3.1 BERT (bert-base-uncased) : 24
4.3.2 RoBERTa (roberta-base) : 25
4.4 Fusion Based 26
4.4.1 Late Fusion 27
4.4.2 Hierarchical Attention Fusion 28
4.4.3 Cross-Modal Transformer Fusion 29
4.4.4 Gated Multimodal Fusion 30
5 Experimentation and Results 32
5.1 Evaluation Metrics 32
5.1.1 Classification Report 32
5.1.2 Confusion Matrix 33
5.2 Text Models 33
5.2.1 BERT(bert-base-uncased) 34
5.2.2 RoBERTa(roberta-base) 34
5.2.3 BiLSTM 34
5.3 Audio Models 35
5.3.1 Bi-directional LSTM 36
5.4 Fusion Models 40
5.4.1 Late Fusion 41
5.4.2 Hierarchical Attention 42
5.4.3 Cross-Modal Transformer 42
5.4.4 Gated Multimodal 42

6 Conclusion and Future Work 46

LIST OF REFERENCES 48

APPENDIX

CHAPTER 1

Introduction

Emotion is a powerful feeling that defines an individual mental state and guides our actions. Researchers have utilized emotion detection and sentiment analysis in the text for various applications, including customer service and chatbots [1], where the study of customer interactions and sentiment in chat dialogues enables improvements in customer care experiences. Emotion detection frameworks have also been used in social media analysis and monitoring[2], allowing for the comprehension of emotions represented in social media postings and comments, as well as insights into public sentiment and opinion. Furthermore, similar frameworks have been used in psychological research and treatment[3], allowing for the assessment of emotional states and development in therapy sessions through the analysis of text or transcriptions[4]. Although recent work on emotion detection for the mentioned applications has been promising, extraction of specific emotions such as fear, nervousness, calmness, confidence, and so on from text and speech to detect liars and deceivers has not been studied enough. Emotion Recognition in Conversations (ERC), which has gained attention in the past few years, is the process of identifying emotions expressed by participants in a conversation. ERC can be highly beneficial in improving human-computer interaction, enhancing customer service, and supporting mental health initiatives, especially in real time.

However, Emotion recognition comes with certain challenges that make it difficult to solve. Emotion can be expressed in different categories and is subjective, varies from person to person. Situational, social, and cultural factors also influence the complex and dynamic nature of emotion. Human evaluation also cannot accurately predict an emotion from a person. Small utterances in a conversation like 'Yeah!!' can have different emotions. The presence of sarcasm, shift of emotion, and context

can happen anytime in a conversation. For example, the IEMOCAP dataset [5] has been annotated by 6 annotators. Yet, there have been different emotion category labeling by different annotators, which explains 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 a deep-learning model but also raise questions about the predicted category of emotion and its accuracy. Identifying emotion through dimensional model[6] can be done and then mapped to the nearest emotion categories. However, the issue arises with the dataset. A dataset with proper dimensional labels is not readily available. While IEMOCAP provides dimensional mappings ranging from 1 to 6, insufficient work has been done on that dataset to fully explore its potential. Extraction of 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 the 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, loudness, and more can be utilized to predict emotion, as they play a crucial role.

In this project, we utilize Deep learning models and Recurrent Neural Networks(RNNs) based on Long short-term memory(LSTM), which can learn context and emotional flow throughout the conversation. LSTMs can capture the dynamic shifts of emotions by remembering earlier utterances. Meanwhile, Simple models like Support Vector Machines (SVM), Naive Bayes, and Decision Trees will serve as the baseline models due to their interpretability and ease of training. 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, Large Language Models (LLMs) like

BERT, GPT-3, or RoBERTa and Transformer models like Audio Spectogram Transformer(AST) and Wav2Vec2 models have demonstrated remarkable performance in various natural language processing tasks.

The remainder of this paper is structured as follows, the subsequent section, Literature Review, provides an overview of existing research in emotion recognition and highlights relevant methodologies and findings. Following that, the Datasets section discusses the available datasets, including their characteristics and limitations. The Methodology section outlines the approaches employed in this study, detailing the feature extraction and model training processes. The Experimentation and Results section presents the experimental setup, including the evaluation metrics, and summarizes the key findings from the experiments, providing insights into the effectiveness of the proposed methods. Finally, the Conclusion and Future section explains possible areas of future research in emotion recognition.

CHAPTER 2

Literature Review

The literature review primarily focuses on different techniques for automated emotion recognition that have been worked on from the beginning, explores existing challenges and approaches adopted for ERC, and finally *Using a categorical model to detect emotion specifically from conversational and dialogue data, investigate different implementations with multimodal data and improve the prediction accuracy.* Survey papers, peer-reviewed journals, research projects, and published papers in this field have been used to address these objectives. Section 2.1 represents an overview of current Emotion modeling and datasets. It explains the different proposed models for emotion and related datasets. Section describes different methods by which emotion can be detected and dive deeper into each technique. Figure 1 shows the organizational structure of the literature review.

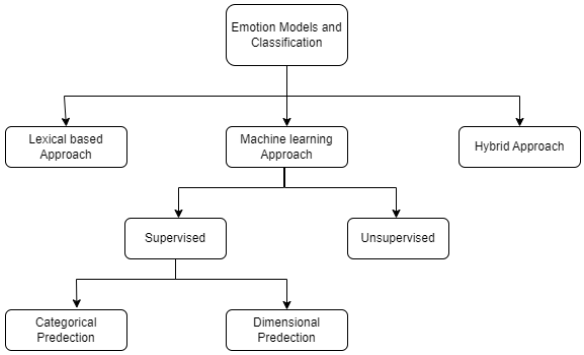

Figure 1: Organizational Structure of Literature Review.

2.1 Emotion Models and Classification

2.1.1 Emotion Models

Emotion is defined generally 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 an individual can only experience one emotion at a time. On the other hand, in the dimensional

model of emotions, the labeling is done on a multi-dimensional space rather than as distinct, separate categories. Ekman et al. [7] proposed 6 basic emotion categories. It is by far the most used and well-known model for Emotion classification. The dimensional model views emotions as positions in a space defined by dimensions like valence (positive/negative), arousal (calm/excited), and dominance (controlled/submissive). Different models like Russell's circumplex, Thayer's energy-stress model, and Plutchik's "emotion wheel" represent emotions within this dimensional space in different ways. Thayer's energy-stress model utilizes the two dimensions of energy and stress. Plutchik wheel of emotions is a circular model for classifying emotions developed by American psychologist Robert Plutchik in 1980. It displays eight primary emotions and their close secondary emotions in 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 us to choose a point in a 2 or 3-dimensional model and gives a wide range of emotions in a point space.

2.1.2 Emotion Datasets

Traditionally, emotion datasets were focused on textual modalities, with most research focusing on 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 data has been labeled, the dimensional annotation by the evaluators are not completely accurate and there is an observable difference. Another dataset designed

for research in emotion recognition, particularly in the context of conversations, is MELD (Multimodal EmotionLines Dataset)[8] - an extension of the EmotionLines dataset featuring all textual, audio, and visual modalities from the Friends TV series. Recently, Google has launched a large-scale dataset, GoEmotions [9] with more fine-grained emotions. Although it's not completely conversational, the dataset features comments extracted from Reddit conversations. The above three datasets focus on multimodal conversational data. Other non-conversational datasets which cover studying basic emotions include EMOV_DB [10] EmoDB [11], RAVEDESS(Ryerson Audio-Visual Database of Emotional Expressions) [12] and CMU[13]. This literature review currently focuses mostly on conversational data IEMOCAP and partly on non-conversation data EMOV to study emotion detection.

2.2 Classification Approaches for Emotion Detection

2.2.1 Lexical & Rule-Based Approach

The Lexical Approach focuses on analyzing text's emotional content by examining words and phrases associated with specific emotions. On the other hand, the Rule-Based Approach relies on predefined sets of rules and linguistic patterns to identify and classify emotions. Combined, they focus on syntactic structures, semantic relationships, and linguistic features to identify the context to classify emotion from the text. [14] & [15] both use lexical & rule-based approach to identify emotion in text for detecting Ekman basic emotions. In [14], the authors identify Keywords, apply rules to exclude unnecessary sentences, and consider negation words to classify the emotion class for the text. Authors in [15] also use a Lexical-based approach with a slightly deeper analysis of textual features. They created a bag of words known as the EmotionWords Set (EWS) for Ekmean standard emotion categories, each word associated with corresponding intensity levels, and then scored the emotion based on its degree. Although this approach uses simple NLP and Syntactical rules to identify emotion labels, such

approaches struggle to classify conversational emotion, especially in utterances with fewer wordings.

2.2.2 Machine Learning

To detect emotions such as fear, nervousness, calmness, confidence, and more from text and speech using machine learning models, a comprehensive and systematic approach will be followed. The literature review has explored various published research and papers. The following sections will explain the review of different implementations for Emotion detection in conversation. Adopting or implementing a Machine Learning model will involve a series of steps depending on the dataset, selecting the model as per the target class or values to be predicted, and feature extraction from the dataset. Firstly, publicly available annotated datasets with target emotions, such as IEMOCAP[5], will be 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 will be performed to represent the emotional content in both text and speech data. For text data, a combination of lexical, syntactic, and semantic features will be considered. Lexical features will involve extracting word-level statistics like term frequencies, term frequency-inverse document frequency (TF-IDF), and n-grams to capture important words and phrases related to emotions. Syntactic features, such as part-of-speech tags, dependency parse trees, and sentiment scores, will provide information about the sentence structure and sentiment expressions. Additionally, semantic features will be derived from pre-trained word embeddings like Word2Vec, GloVe, or FastText to capture contextual meanings.

For speech data, acoustic features will be extracted to capture the prosodic and spectral characteristics related to emotions. Popular acoustic features like Mel-

frequency cepstral coefficients (MFCCs), pitch, energy, and formants will be computed from the speech signals. These features will be analyzed over small windows of audio to capture temporal variations in emotional expression. Once the relevant features are extracted, different machine learning models will be explored, including both simple and deep learning approaches. Simple models like Support Vector Machines (SVM), Naive Bayes, and Decision Trees will 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 will be explored for their ability to capture complex patterns and dependencies in the data. In addition, Large Language Models (LLMs) like BERT, GPT-3, or RoBERTa, demonstrated better performance in NLP tasks, can also be incorporated. These models can be fine-tuned using the emotional datasets to specialize in detecting emotions from the text. During the model training process, hyperparameter tuning and cross-validation will be implemented to make model reliable. The best-performing models will be selected based on evaluation metrics like accuracy, precision, recall, F1-score, and confusion matrix results. The most informative features will be identified using techniques like Recursive Feature Elimination (RFE) or feature importance analysis from tree-based models.

2.2.2.1 Recurrence-based Models

Traditional methods often struggle to capture the dialog context and dynamics of conversation, where emotions can shift subtly with each turn of phrase. This is where recurrence-based models, particularly Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Gated Recurrent Unit(GRUs) offer a deeper understanding of emotional expression in dialogue by capturing context and temporal variations of the conversation. RNNs excel at capturing the dependencies between

the between words that shape emotional meaning.

LSTM & BiLSTM: Recurrence-based models, such as Long Short-Term Memory (LSTM) networks and Bidirectional LSTMs (BiLSTM), 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.LSTMs are a type of recurrent neural network (RNN), designed to address the problem of vanishing gradient associated with traditional RNNs and excel at capturing sequential dependencies and context over extended distances. On the other hand, BiLSTMs enhance the capabilities of LSTMs by processing sentence in both directions. This approach allows the model to capture not only past but also future context for each time step, providing a more comprehensive understanding of the conversation.

[16, 17, 18] worked on the dataset provided by International Workshop on Semantic Evaluation(SemEval-2019) task 3, which consists of 15k tweet conversational pairs. [16] opted for the Glove twitter.3B300 embedding on LSTM, Naive Bayes, and SVM classifiers, to detect emotions in textual conversations, and reports an accuracy of 85%. On the same dataset [17] experiments with 3 embedding approaches (Word2Vec, FastText, and Glove) on a BiLSTM network. The author achieves the highest F1 score of 69.63% with Glove Embeddings. On the other hand, [18] applies offensive filtering, PII filtering, and language filtering on the dialog to ensure quality and relevance. Authors use a combination of lexical features neural models, ensembling architectures with fine-tuned BERT, and LSTM. [19] compares all the models used on the dataset and concludes that BiLSTMs/LSTMs were the most frequent models, and GRU, CNN models were used in other implementations. Attention mechanism, ensemble approaches, and Transfer learning using BERT, ELMo, and ULMFit were popular among most implementations.

SemEval dataset is a textual conversational pair with 3-4 dialogs. For Emotion Recognition in conversations, a more detailed conversational dataset is necessary, like MELD and IEMOCAP. [20] propose a novel method combining the RoBERTa model (a robustly optimized BERT pretraining approach) with a BiLSTM network for Emotion Recognition in conversations on the MELD textual dataset. The model achieved a weighted average F1 score of 60.12%. [21] takes it a different approach and detects Emotion from speech using two datasets, namely MSP-Podcast and IEMOCAP. Implements a listener Adaptive (LA) model to address listener-dependent emotion perception. It consists of an encoder, a decoder, listener embedding layer, Adaptation layers like Adaptive Fully-Connected (AFC), Adaptive LSTM (ALSTM), and Adaptive CNN (ACNN). The author achieves a slightly higher accuracy of 63.2% compared to [20].

All the above implementations and research perform categorical classification for detecting emotion. Although that can work few class labels, as the number of classes increases, there would be a negative effect on accuracy. Authors in [22, 23] instead of predicting the category of emotion class, work on the dimension prediction.. [22] introduces a deep neural network model designed to track continuous changes in emotion, particularly in terms of arousal and valence. The architecture includes convolutional neural network (CNN) BiLSTM layers to handle temporal and spectral variations on the SEMAINE and RECOLA databases. It achieved a concordance correlation coefficient (CCC) of 0.680 and 0.506 on Arousal and valence respectively. As the performance of valence prediction is being affected in dimensional space, [23] proposes a method that integrates acoustic features with text features, converting words into vectors. This novel approach significantly improved the prediction accuracy for valence in both single-task learning. Continuous dimension prediction of Arousal,

valence and dominance could describe a wide range of emotion results with intensity.

GRU: Gated Recurrent Units (GRUs) are a type of recurrent neural network (RNN) architecture used in deep learning. GRUs 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. [24] focuses on sentiment and emotion detection in conversation using multimodal data(IEMOCAP, CMU-MOSE). The author proposes Recurrent neural network (RNN)-based method(GRU) and 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.[25] utilizes GRU to monitor the speaker's evolving mental state through the conversation and [26] feed the input features to GRU for extracting the global contextual information.. [27] proposes a Novel architecture with an autoencoder on IEMOCAP and EMODB datasets to extract deep emotion features from speech. The auto encoder consists of convolution parts, instance normalization, dropout layers, and a Gated Recurrent Unit (GRU).

CNN & GCN: CNNs can be a powerful tool for emotion detection in conversations by understanding the context and capturing sequential patterns from large datasets. 1D CNNs, Recurrent CNNs (RCNNs) and Graph CNNs(GCNs) will offer unique advantages in extracting features and predicting the target emotion. [28] delves into the detection of multiple emotions in text data. The authors highlight the presence of multiple emotions within conversational data and introduce a custom-built Convolutional Neural Network (CNN) architecture for this purpose. 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. Combination of CNN with GloVe embedding outperforms other combinations with a Jaccard Index of 0.6463.

[29] construct an utterance-level Graph Convolutional Network (U-GCN) focusing on semantic correlations among utterances, and speaker-level GCN (S-GCN) capturing correlations between new utterances and speaker emotions. Features are extracted using RoBERTa. I-GCN model outperforms baseline models with an F1 Score of 65.4%. [30] does a comparative study of CNNs and Convolutional Recurrent Neural Networks(CRNNs) for emotion recognition in speech. Two kinds of speech features MFCC and GFCC are used as the Input to Neural Network. CRNN uses additional recurrent layer(GRU) to extract the temporal Information. After training the model, it is observed that CRNN has higher fitting degree and accuracy compared to CNN. CRNN has performed with testing accuracy of 77.20% on MFCC features and 73.40% on GFCC features. [31] introduces Dialogue Graph Convolutional Network (DialogueGCN), a graph-based approach for emotion recognition in conversations (ERC). This method overcomes limitations of RNN-based models by using a directed graph to represent conversations, effectively capturing both sequential and speaker-level contexts. Evaluated on benchmark datasets like IEMOCAP, AVEC, and MELD, DialogueGCN achieved an accuracy of 64.18%.

In the case of dimensional space, [32] introduces a three-layer model with fuzzy inference systems for estimating these dimensions, utilizing speech features from prosodic, spectral, and glottal waveform sources. It achieves a smaller mean absolute error and higher correlation with human evaluators. [33] proposes 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 concordance correlation coefficient (CCC)

of 0.7635, 0.2894, and 0.7130 on Arousal, valence, and Dominance respectively.

2.2.2.2 Other Models

Along with Deep Learning like LSTMs and CNNs, which have gained prominence in emotion detection from conversations, simple machine learning (Baseline) models still hold value under certain circumstances especially when the data is simple text instead of conversations. Along with NLP, authors in [15] use standard classifiers like SMO and J48 for text (tweets) classification and [34] implements classical Machine Learning models like Naive Bayes (NB) and k-nearest neighbor algorithms (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 won't perform well on conversational or multi-modal data. Taking a different approach [26] proposes Emotion Detec- tion Reinforcement Learning Framework (EDRLF) to detect emotions in conversations by considering both the influence of preceding emotional states (ES) and contextual information from previous utterances on MELD dataset. The authors extracts textual and acoustic features separately from the dataset utilizing GRUs and combine it with a reinforcement learning agent (D-Q network) to make sequential emotion detection decision. EDRLF gave highest w-average F1 of 60.2 on multi-modal. This will be highly effective as an incremental learning model.

2.2.3 Hybrid Approach

The hybrid approach follows a combination of both Lexical and Machine Learning. [15] along with Emotion-Word set, combine it with SMO and J48 classifiers to achieve an accuracy of 91.7% and 85.4% respectively. [35] also proposes a hybrid approach for the detection of multiple emotions in text and conversations on the combination of 4 datasets ISEARs, MELD, EmoDB, and GoEmotions. The hybrid model consists of manually written rules and also uses a pre-trained model Sentence-BERT to get the

Table 1: Analysis of Models with Categorical Prediction

Paper	Dataset	ML Model	Modal	Metrics	Classes	Results
[16]	SemEval	GloVe+LSTM	Text	Accuracy	5	85
[17]	SemEval	(Word2Vec,FastText)+BiLSTM	Text	F1 Score	4	69.63
[18]	SemEval	Embeddings+BiLSTM	Text	Micro F1	3	75.8
[20]	MELD	RoBERTa+BiLSTM	Text	F1 Score	6	60.12
[21]	IEMOCAP	ALSTM & ACNN	Audio	W.Accuracy	4	61.6
[25]	IEMOCAP	VGG+GRU	Multi(3)	F1-Score	5	65.4
[27]	IEMOCAP	CNN+GRU	Audio	UnW.Accuracy	4	71.2
[24]	IEMOCAP	(GloVe,openSmile)+GRU	Multi	F1-Score	5	73.3
[31]	IEMOCAP	CNN+GCN	A+V	F1-Score	6	64.18
[29]	IEMOCAP	RoBERTa+GCN	A+V	F1 Score	6	65.4
[36]	IEMOCAP	LSTM-Attn	T	F1 Score	4	63.3
[36]	IEMOCAP	AttnFusion	A+T	W.Accuracy	4	70.4
[37]	IEMOCAP	MFGCN	A+T	W.Accuracy	4	78.3
[38]	IEMOCAP	Attn+CNN	A+T	F1 Score	4	66.1
[39]	IEMOCAP	Transformer-Fusion	A+T+V	F1 Score	4	84.1
[39]	MELD	Transformer-Fusion	A+T+V	F1 Score	7	63.9
[40]	IEMOCAP	Gated-Fusion	A+T+V	W.Accuracy	4	72.39

best emotions for each dialogue and generate multiple emotion tags. The addition of Machine learning to Lexical approaches can help achieve better results in conversations as compared to only the Lexical approach as in 2.2.1

Table 2: Analysis of Models with Dimensional Prediction

Paper	Dataset	ML	Modal	Metrics	Results
Yang et al. [22]	SEMAINE	CNN+BiLSTM	Speech	CCC	(A,V)=0.68,0.50
Dim. et al. [23]	IEMOCAP	LSTM	Multi-2	Mean CCC	0.48
Li et al. [32]	Fujistu(1),Berlin(2)	3-Layer	Audio	MAE	(.16,.12)(1),(.37,.18)(2)

CHAPTER 3

Dataset

Two primary datasets were utilized: IEMOCAP[5] and EMOV[10]. This project focuses on conversational emotion detection, primarily utilizing the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset and EMOV is used as an additional dataset for speech emotion recognition.

3.1 IEMOCAP

The dataset consists of approximately 12 hours of audio-visual data of 10 actors (5 male and 5 female) who were recorded while performing scripted dialogues with wide range of emotions including happiness, sadness, anger, frustration, surprise, and neutral. The dataset includes both audio and visual data, as well as transcriptions of the dialogues. The annotations include both discrete emotion labels (e.g., "happy", "sad", etc.) and dimensional emotion values like arousal, valence and dominance.

For this project we are currently focusing only on textual and audio data for predicting categorical labels. The dataset includes annotations of the emotions expressed by the actors at each point in the dialogues by a total of 6 evaluators. Each dialogue is labeled by 3 evaluators. These are spread across 5 session and a total of 150 conversations and 9953 dialogues.

Figure 2 and Figure 3 shows the spread of total dialogues(samples) and conversations across 5 sessions and Figure 4 shows the distribution of dimensional values of dataset. The emotion categories will be discussed in preprocessing section of Chapter 4.

There are few challenges associated with using the IEMOCAP dataset for emotion recognition in conversations. One challenge is the the size of the dataset, 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 mismatching when evaluated by more than one which indicates the complexity of labeling categories for conversational data. To address these challenges, we adopted few preprocessing steps for sample selection, feature extraction and model training.

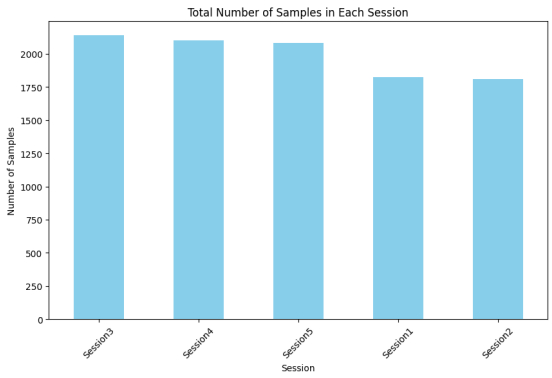

Figure 2: Session wise spread of IEMOCAP dataset

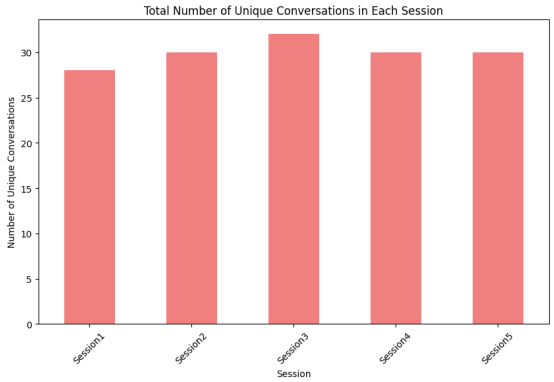

Figure 3: Conversation wise spread of IEMOCAP datatset

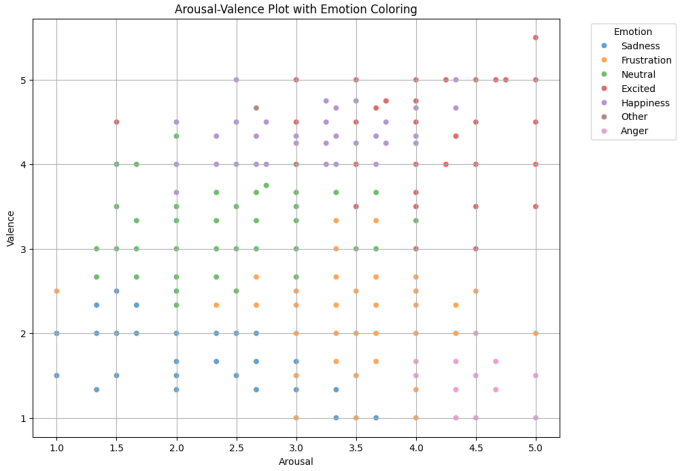

Figure 4: Arousal and Valence distribution across dataset

3.2 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 amused. The dataset includes a total

of 6,893 audio files, with varying durations ranging from 1 to 10 seconds. Figure 5 shows the distribution of emotion categories across dataset.

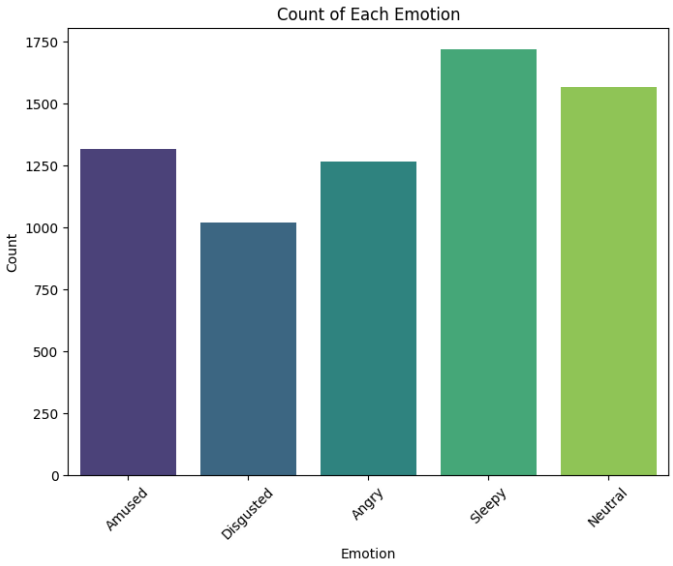

Figure 5: Arousal and Valence distribution across dataset

CHAPTER 4

Methodology

For this research, we primarily focus on textual and audio modals for emotion detection. Separate models are trained on textual transcripts, audio and fusion of both text and audio datasets with key preprocessing steps and techniques for each modal.

4.1 Dataset Preparation

IEMOCAP dataset is preprocessed to gather all the textual transcripts and audio dialogues from each conversation. This final dataframe includes information such as the 3 emotion labels assigned by annotators, the time frame of each dialogue, audio dialogue, speaker IDs, session IDs, and conversation ID for each sample. Since, each dialogue has been labelled by 3 different annotators, we select the samples where the annotated categorical emotion value matched by atleast 2 annotators. The final dataset contains 7766 samples. The distribution of samples and categories across the dataset is shown in Figure 6 and Figure 7.

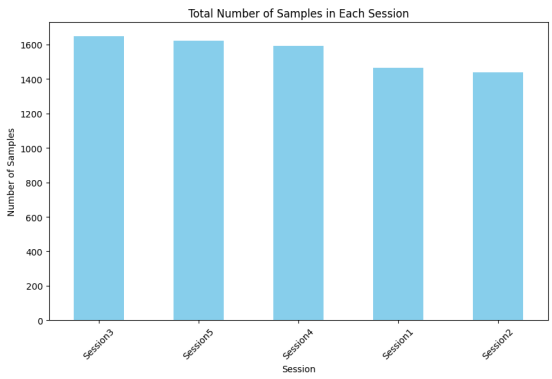

Figure 6: Sample wise spread after selection

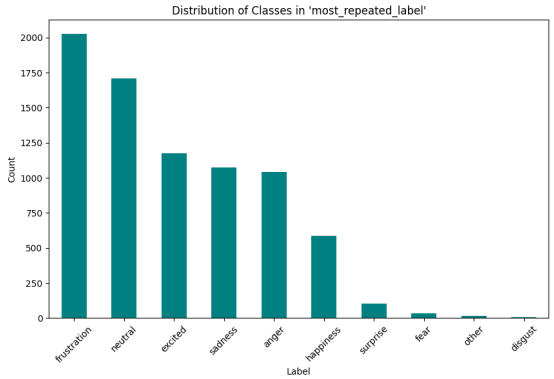

Figure 7: Category wise spread of IEMOCAP datatset

The IEMOCAP dataset provides the textual transcripts for each conversation. This technique aims to classify emotions using a deep learning model with text data.

We utilize multiple transformer-based models like DistilBERT, RoBERTa, and BERT which brings unique features to extract contextual embeddings.

4.1.1 Data Collection and Preprocessing

The dataset was preprocessed as follows:

1. **Label Selection and Mapping**: Initially, the dataset has emotion labeled in with 10 categories. Labels like surprise,fear, other and disgust have been dropped as the class size is 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.''
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, short and ambiguous sentences were not considered.
4. **Label Encoding**: Labels were converted to a numerical format using `LabelEncoder` from scikit-learn, which enabled compatibility with the model's output layer.

Figure 8 shows the distribution of samples across 4 categories classes after Preprocessing.

4.1.2 Acoustic Feature Extraction

In this method, we focus on extracting a comprehensive set of acoustic features from the audio conversations from IEMOCAP and EMOV, then train the model with the extracted features. These features include:

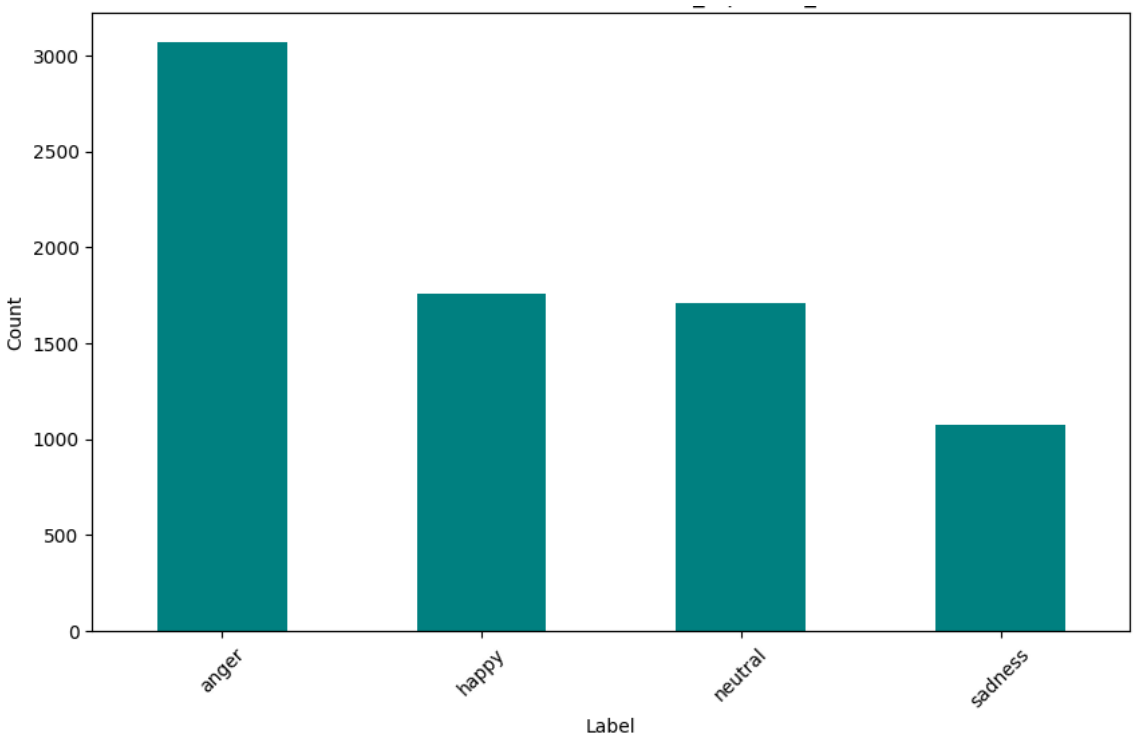

Figure 8: Category wise spread of IEMOCAP dataset

- **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 the sound and is essential for detecting tonal variations in speech. We compute the mean and standard deviation of the pitch values.
- **Energy**: The root mean square energy (RMSE) measures the amplitude of the audio signal 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 is a measure of 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 rolloff.
- **Spectral Bandwidth**: This feature measures the width of the spectral band and helps in identifying the spread of the spectrum. We calculate the mean spectral bandwidth.
- **Chroma Features**: These features represent the 12 different pitch classes and provide a harmonic representation of the audio signal. 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, providing insights into the noisiness and percussiveness of the sound. We compute both the mean and standard deviation of ZCR.

The extracted features with and without Noise filter and then used to train with various deep learning model. Models like Bi-directional LSTM, Transformer, and a simple neural network are used, all implemented using the Keras library in Python. Below is a detailed description of each model, including their respective architectures and hyperparameters.These models are capable of capturing temporal dependencies in the audio data, making it suitable for this task. Details of each models are as follows:

4.2 Audio Based

Features extracted from section 4.2 are trained with the below model.

4.2.1 Bi-directional LSTM Model :

Features extracted from section 4.2 are trained with the Bi-directional LSTM model.The Bi-directional LSTM model consists of two Bi-directional LSTM layers of

with 128 and 64 units, respectively. LSTM is a Recurrent Neural Network used to capture sequential data dependencies and reducing problems like vanishing gradients in traditional RNNs. The cells maintain information across long sequences and gates are used to control the flow maintaining the focus on context.

Dropout layers with a rate of 0.2 are added after each layer to prevent overfitting. This model includes Adam optimizer with a learning rate of 0.001, categorical cross-entropy as the loss function, and accuracy as the evaluation metric. The Keras-Tuner searches for optimal configurations of LSTM units and dropout rates. Attention is used to enhance the focus on significant parts of the sequence.

The steps involved for training the above model are as follows:

- **Data Preprocessing**: he audio files were loaded with a sample rate of 16 kHz and converted to mono channel using the Librosa library to extract relevant segments and convert them into a suitable format for feature extraction.
- **Feature Extraction**: The above mentioned acoustic features are extracted using the Librosa library.
- **Model Training**: The extracted features are used to train a 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.

4.2.2 Audio Spectrogram Transformer (AST)

The Audio Spectrogram Transformer (AST) [41] 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.

Process:

1. **Spectrogram Generation**: The raw audio waveform is transformed into a spectrogram.

2. **Patch Embeddings**: The spectrogram is divided into smaller patches, which are then embedded into a lower-dimensional space.

3. **Transformer Encoder**: These embedded patches are fed into a Transformer encoder, which processes the sequence using self-attention to capture long-range dependencies.

4. **Classification Head**: The output from the Transformer encoder is passed through a classification head to predict emotion labels.

Figure 9 shows the architecture of the AST model used in the experimentation.

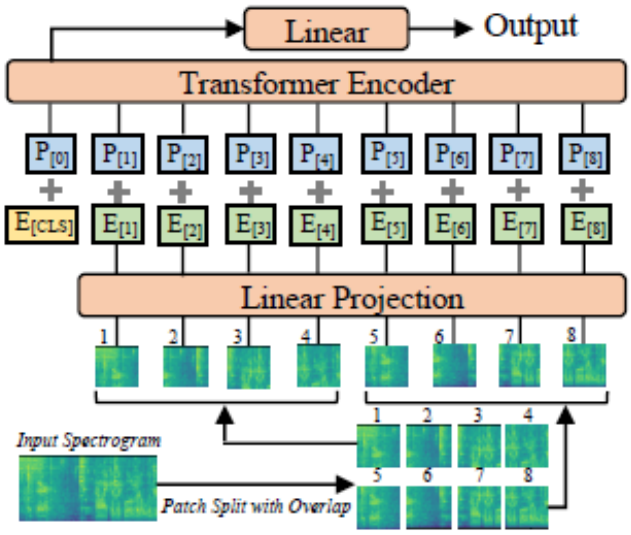

Figure 9: The architecture of the AST [41].

4.2.3 Wav2Vec2

Wav2Vec2 [42] 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 model follows the below steps

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

Figure 10 shows the architecture of the Wav2Vec2 model used in the experimentation.

4.3 Text Based

Model Architecture The model used in this study leverages a pre-trained transformer models for feature extraction, followed by a classification layer which are well-suited for capturing contextual information in the text. Transformer based models are chosen, primarily due to their ability to capture contextual information and semantic relationships with the help of attention mechanism [43]. 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 mechanism.

4.3.1 BERT (bert-base-uncased) :

Introduced by Google, The BERT model is well known for capturing bidirectional context in language, is used to achieve fine-grained emotion detection. BERT is

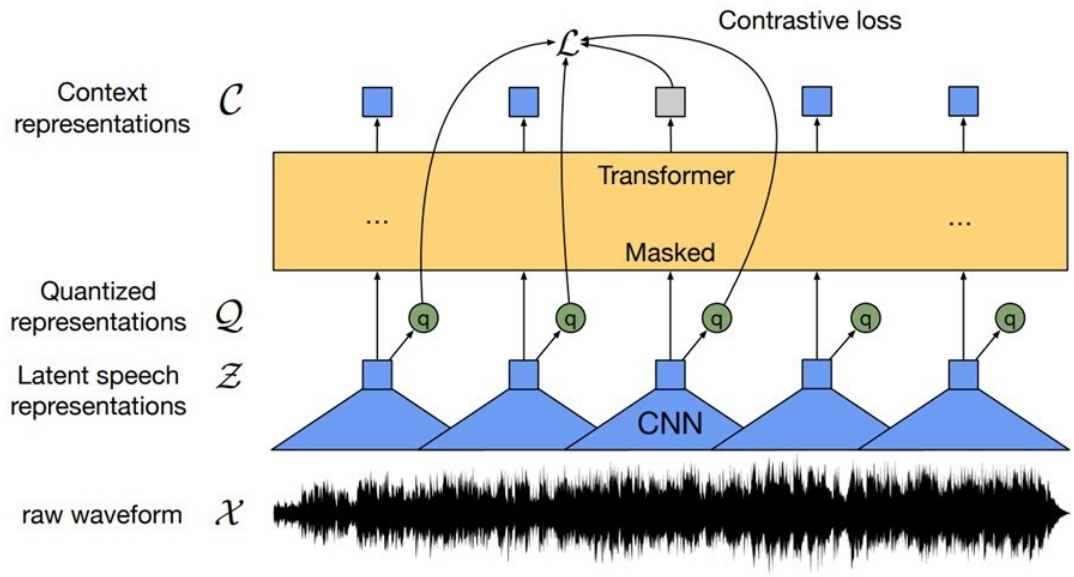

Figure 10: The architecture of the Wav2Vec2 [42].

pretrained on two tasks. MLM(Masked Language Modeling) which predicts the masked words in a sentence and Next sentence prediction.

4.3.2 RoBERTa (roberta-base) :

RoBERTa, a robustly optimized version of BERT developed by Facebook provides contextual embeddings and captures larger language patterns in emotions effectively. RoBERTa removes the NSP, focusing only on MLM for better token representations.

Both the encoders are trained on two classifiers:

1. A fully connected layer was added at the end to map hidden states to the number of target classes (emotions).

2. Bi-directional LSTM model which processes the data sequence in both directions

- The models were fine-tuned on the emotion classification task with a small learning rate to avoid disrupting pre-trained weights.

- **Optimizer and Loss Function**: The AdamW optimizer was used with a learning rate of 2×10^{-5} for all models, providing stable convergence. Cross-Entropy loss was used to compute the error during training.

Data Splitting and Training The data was split into an 80:20 ratio for training and testing. For each model, the text samples were tokenized using the respective tokenizers with padding and truncation to a maximum length of 128 tokens. The text was then converted into PyTorch tensors for compatibility with the models.

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

4.4 Fusion Based

Fusion based models combine multiple modalities, to form a multimodal understanding. For IEMOCAP, features from audio and text are combined and at a certain stage in the model pipeline. The stage at which we combine the features decides the kind of fusion. Pre-trained Roberta model is used to extracts text embeddings which will be used text features, similar to Section 4.3. Features extracted from Section

4.2 are used as audio features. Both these features are combined to predit the target categorical emotion.

In this report, 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 combine text and audio features in unique way for emotion classification.

4.4.1 Late Fusion

The Late Fusion model is one of the most popular form of fusion model. In this technique we processes text and audio independently and combines their outputs at a later stage to make the final prediction.The independent processing of each modlaity allows it to specialize, and their logits are averaged to form the final classification. This is a simple and effective technique, especially the modalities are loosely coupled, but has limited interaction between modalities. Late fusion incorporated follows the below steps:

- **Text Processing:**: Contextual embeddings are extracted from textual transcripts by using pretrained RoBERTa transformer and a dedicated text classifier is implemented as a feed-forward neural network to processes 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 text classifier and the audio classifier are combined at the logit level using a simple averaging mechanism. Then, the combined logits are then used to predict the final class probabilities

Figure 11 shows the architecture of the model used in our experimentation.

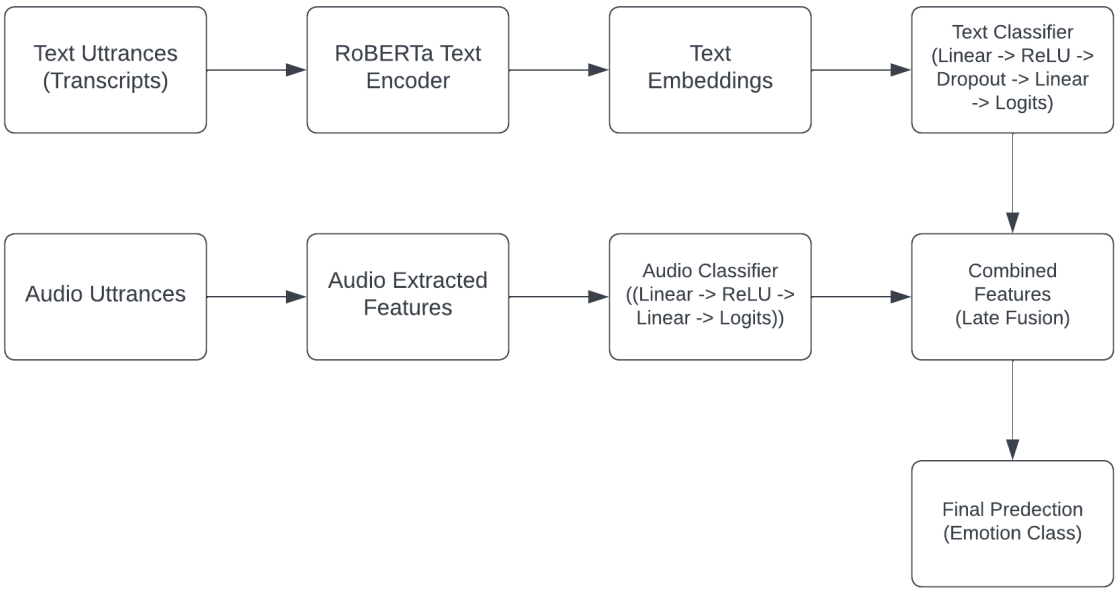

Figure 11: The architecture of the model used for Late fusion.

4.4.2 Hierarchical Attention Fusion

This 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 suited for tasks where specific text phrases or audio signals dominate. Hierarchical Attention Fusion incorporated follows the below steps:

- **Text Processing:**: Contextual embeddings are extracted from textual transcripts by using 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 similarly processed with 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.

Figure 12 shows the architecture of the Hierarchical Attention Fusion used in our experimentation.

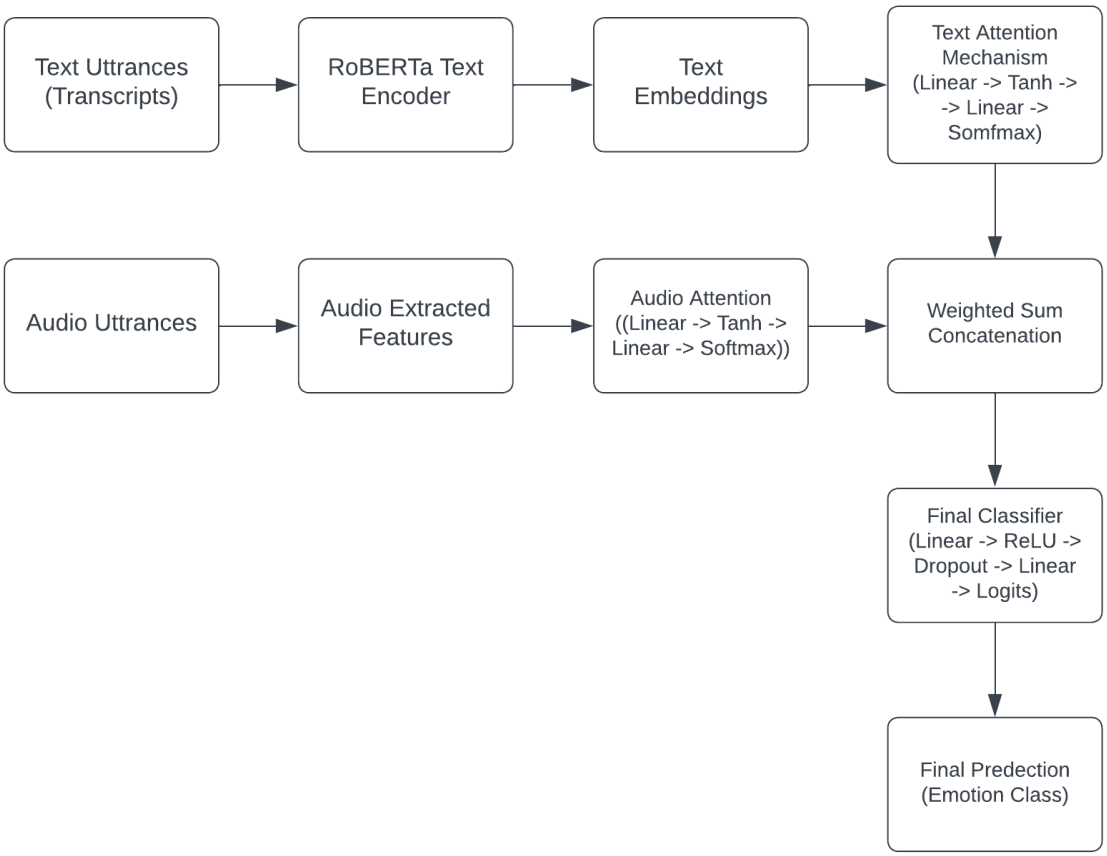

Figure 12: The architecture of Hierarchical Attention Fusion.

4.4.3 Cross-Modal Transformer Fusion

The Cross-Modal Transformer Fusion uses a transformer encoder to fuse text and audio features. The model combines embeddings from both modalities with self-attention layers to capture complex relationships. This architecture specalizes in capturing long-range dependencies and interactions. The model folows the below steps:

- **Text Processing:**: Contextual embeddings are extracted from textual transcripts by using 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 8 attention heads and 2 layers is used to model interactions between text and audio features. The pooled features are fused and classified into emotion categories.

Figure 13 shows the architecture of the Cross-Modal Transformer Fusion used in our experimentation.

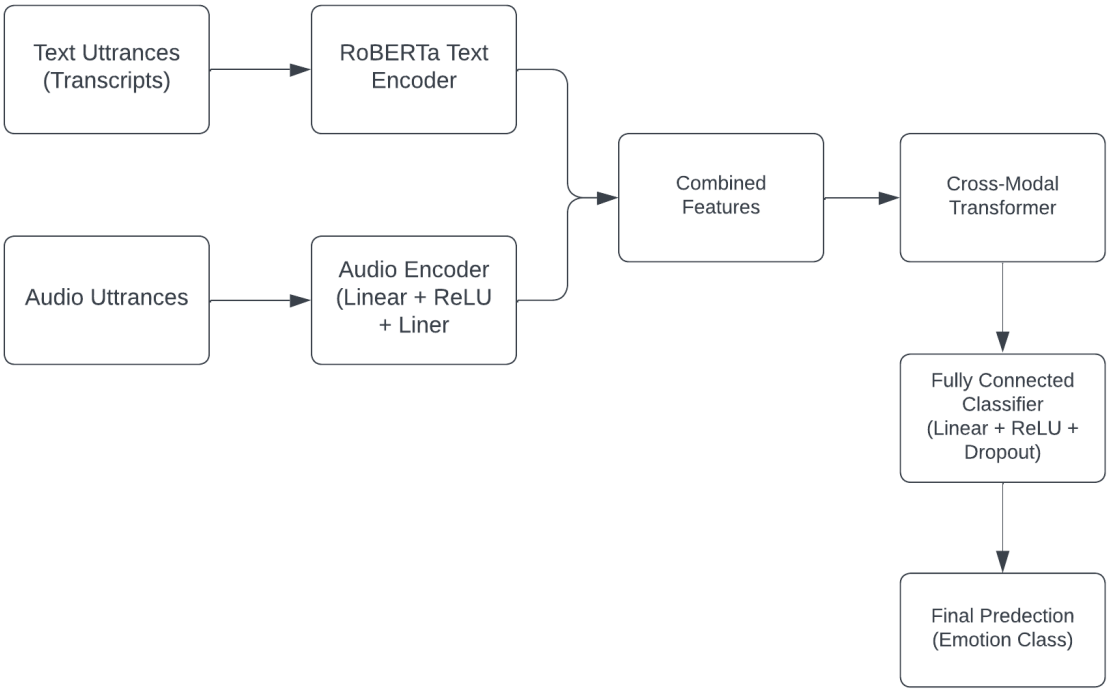

Figure 13: The architecture of the used for Cross-Modal Transformer fusion.

4.4.4 Gated Multimodal Fusion

Gated Multimodal Fusion employs gating mechanisms to assign weights to text and audio features dynamically. The gates decide the contribution of each modality

to the final prediction. This mechanism ensures robust integration of modalities, even when one is noisy or unreliable.

The model folows the below steps:

- **Text Processing:**: Contextual embeddings are extracted from textual transcripts by using 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.

Figure 14 shows the architecture of Gated Multimodal Fusion.

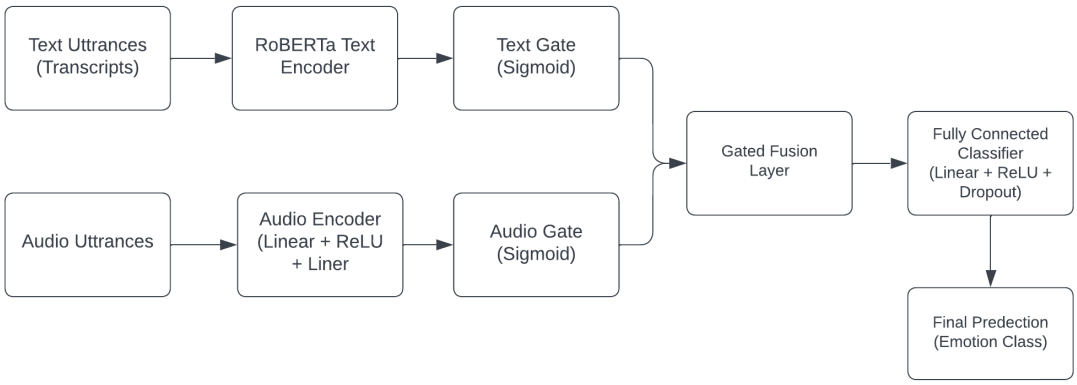

Figure 14: The architecture of the used for Gated MultiModal fusion.

CHAPTER 5

Experimentation and Results

In this section, we present the results of our experimentation with various models for the classification task. The experimentation try to improve classification scores of Emotion classification compared to baseline scores mentioned in table 1.

5.1 Evaluation Metrics

To evaluate the performance, we generated Confusion Matrix and the Classification Report for each model, which includes precision, recall, and F1-score. These metrics are crucial for calculation performance across different emotion classes, especially in a multi-class classification task. They are then compared with the baseline models from Literature review and Table 1

For each model, the evaluation was performed on the test set after training the model, and the following metrics were computed:

5.1.1 Classification Report

The Classification Report provides key performance metrics for each emotion class in the dataset which include:

- **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 correctly identify all positive instances. 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**: The weighted average F1-score accounts for the support (number of true instances) of each class. 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.

  The formula for calculating the weighted average F1-score is:

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

Since IEMOCAP dataset is not balanced, we used Weighted average F1-score as primary metric for evaluation.

The models were evaluated on IEMOCAP dataset with four emotional categories: "happy", "sad", "neutral", and "angry". The weighted average F1-score is particularly useful in situations where the classes are imbalanced, ensuring that the performance on each class is appropriately reflected in the final score.

5.1.2 Confusion Matrix

Confusion matrix is a table used to evaluate the performance of a classification model. It compares the predicted labels with the actual labels and claculates, displays metrics like accuracy, precision, recall, and F1-score

5.2 Text Models

In this study, pre-trained transformer models RoBERTa, and BERT are utilized, along with BiLATM model for comparision. Each of these models was fine-tuned on the IEMOCAP dataset for emotion detection tasks.

5.2.1 BERT(bert-base-uncased)

BERT is known for its ability to capture bidirectional context in a language. BERT achieved excellent performance compared to many baseline models for Emotion classification in text. The model is trained with below hyperparameters and achieved Weighted F1-SCore of 69.92.

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

5.2.2 RoBERTa(roberta-base)

RoBERTa - robustly optimized version of BERT, is known for capturing contextual information and offering improved performance. RoBERTa has performed with Weighted F1-SCore of 72.37. Figure 15 shows the Confusion matrix of RoBERTa based classifcation. The model is trained with below hyperparameters

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

5.2.3 BiLSTM

BiLSTM (Bidirectional Long Short-Term Memory) networks are widely used for sequential data processing tasks due to their ability to capture long-term dependencies in both directions. BiLSTM model was trained with the following hyperparameters and evaluated using 5-fold cross-validation: RoBERTa with BiLSTM has outperformed

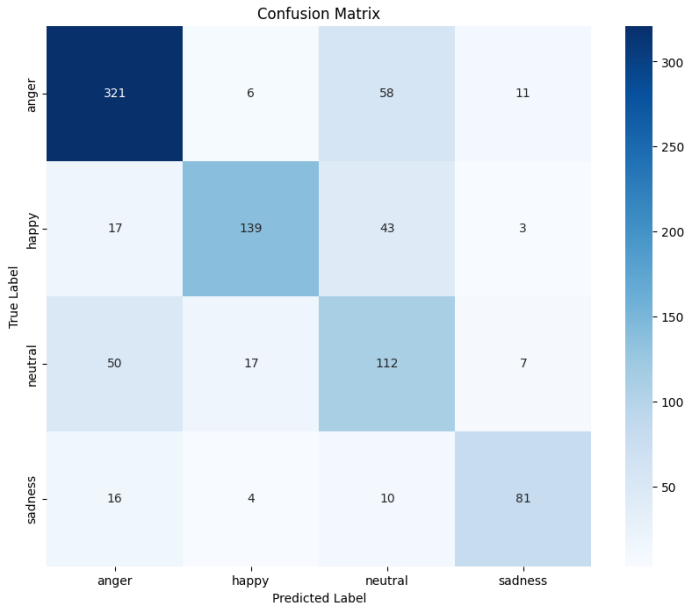

Figure 15: Confusion Matrix for RoBERTa based Text modal Emotion Classification.

all the other models in Emotion Classification with Weighted F1-SCore of 73.33 and BERT with BiLSTM achieved Weighted F1-SCore of 72.75. The models were trained with below hyperparameters.

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

Figure 16 and 17 shows the confusion matrix obtained of BiLSTM classification models on test data with RoBERTa and BERT respectively.

5.3 Audio Models

Models were trained with Acoustic features on BiLSTM, AST, and Wav2Vec2 uses its own encoders for extracting features. For all these three models, we are considering minimum token size of 10 for a sample utterance similar to Text based.

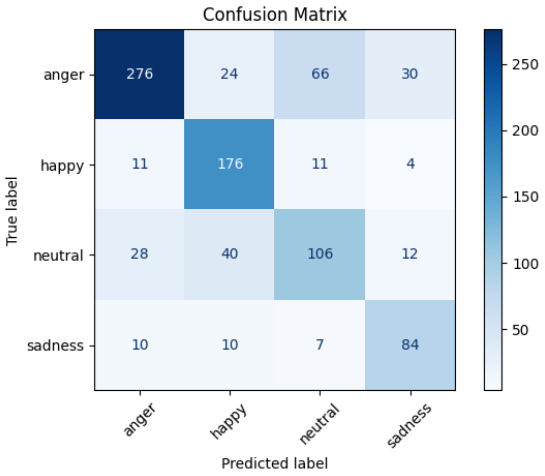

Figure 16: RoBERTa with BiLSTM.

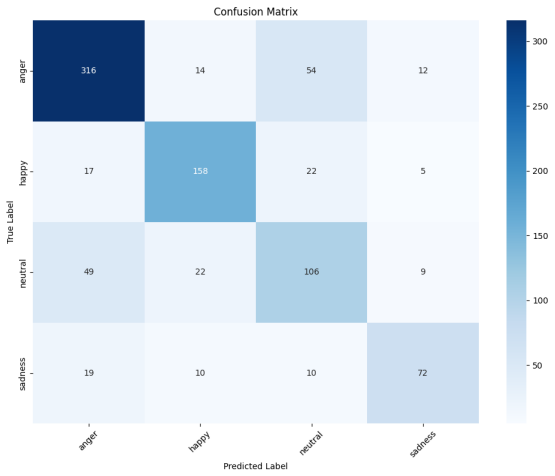

Figure 17: BERT with BiLSTM.

There were total of 4471 utterances which has token size of 10.

5.3.1 Bi-directional LSTM

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

- **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

The model achieved an F1 score of 62.08% on IEMOCAP and 97% on EMOV

datasets respectively. Figure 18 and 19 shows the confusion matrix of BiLSTM on IEMOCAP and EMOV datasets respectively

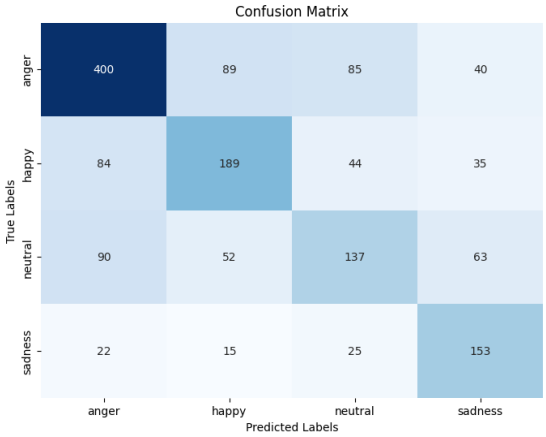

Figure 18: IEMOCAP with BiLSTM.

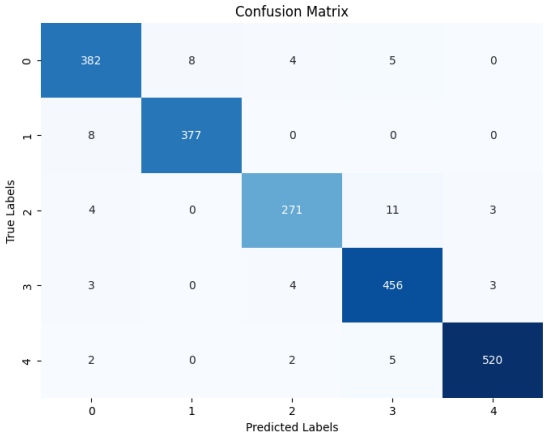

Figure 19: EMOV with BiLSTM.

5.3.1.1 Audio Spectrogram Transformer (AST)

The AST model was fine-tuned on 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 80:20 ratio. Audio features were extracted using the AST feature extractor and audio clips were truncated or padded to a maximum length of 10 seconds

Model Architecture and Initialization

- Pre-trained model: `MIT/ast-finetuned-audioset-10-10-0.4593`.
- Number of trainable parameters: Approximately 5.3 million.
- Output layer: Configured for N-class emotion classification with a `softmax` activation function.

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.

The model achieved an F1 score of 64.99% on IEMOCAP, which is better compared to BiLSTM which was trained on Acoustic features and 98.96% on EMOV dataset. Figure 20 and 21 shows the confusion matrix of AST on IEMOCAP and EMOV datasets respectively.

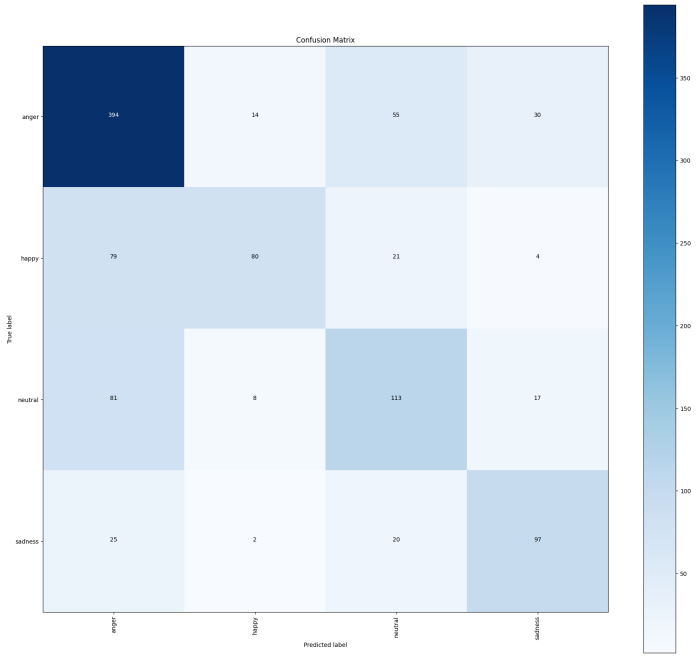

Figure 20: Confusion Matrix of AST with IEMOCAP.

5.3.1.2 Wav2Vec2 Model

The Wav2Vec2 model was fine-tuned on the IEMOCAP dataset using the pre-trained `facebook/wav2vec2-base-960h` model. The dataset was split into training and testing sets with 80:20 ratio. Audio features were extracted using the Wav2Vec2 feature extractor and audio clips were truncated or padded to a maximum length of

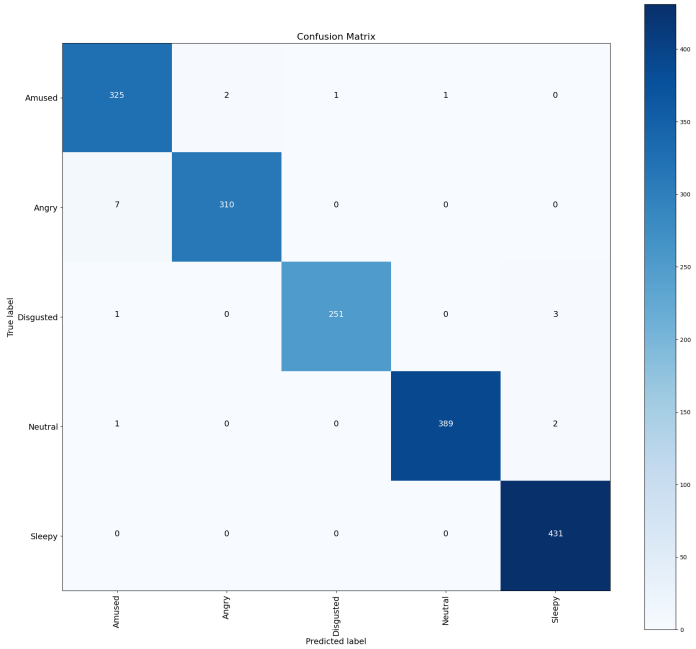

Figure 21: Confusion Matrix of AST with EMOV.

10 seconds

Model Architecture and Initialization

- Pre-trained model: `facebook/wav2vec2-base-960h`.
- Number of trainable parameters: Approximately 95 million.
- Output layer: Configured for N-class emotion classification with a `softmax` activation function.

Training Configuration

- **Batch Size:** 16
- **Number of Epochs:** 30
- **Learning Rate:** 1×10^{-5}
- **Warmup 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.

The model achieved Weighted F1 score of 66.81% which is better compared to both the above audio based models on IEMOCAP and 99.48% on EMOV dataset. Figure 22 and 23 shows the confusion matrix of Wav2Vec2 on IEMOCAP and EMOV datasets respectively.

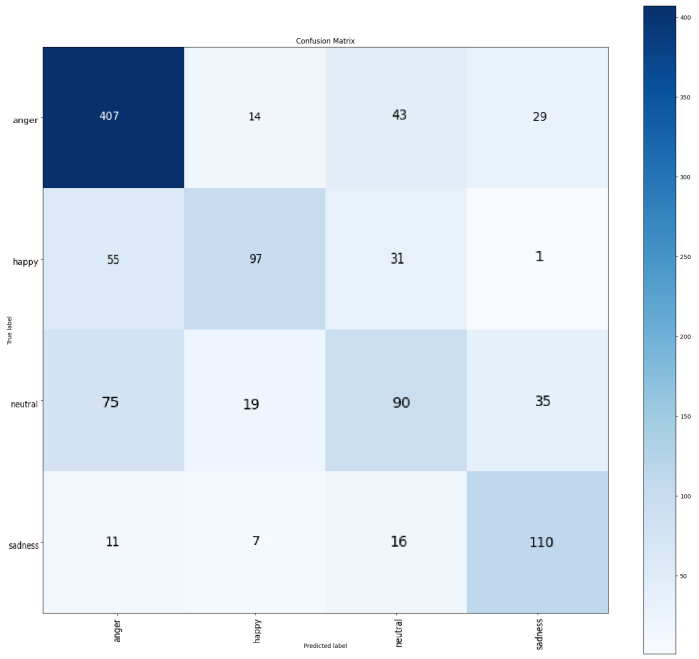

Figure 22: Confusion Matrix of Wav2Vec2 on IEMOCAP.

5.4 Fusion Models

Fusion models performed the best compared to other models in Text and Audio Modal. For this audio features were 81 dimensions Acoustic features extracted as explained in Section 4.2. Four fusion-based models were implemented and compared, which are Late Fusion, Hierarchical Attention, Cross-Modal Transformer, and Gated Multimodal Fusion. Each model was trained with following training configurations.

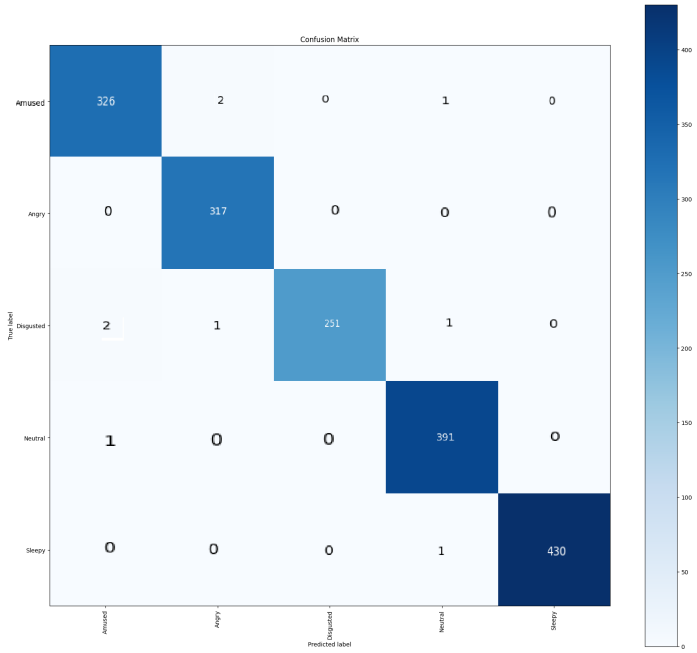

Figure 23: Confusion Matrix of Wav2Vec2 on EMOV.

Training Configuration

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

Figure 24 shows the confusion matrix of Late Fusion model on IEMOCAP.

5.4.1 Late Fusion

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

The Late Fusion model achieved a test weightrf F1 score of **76.62%**. Figure 24 shows the confusion matrix.

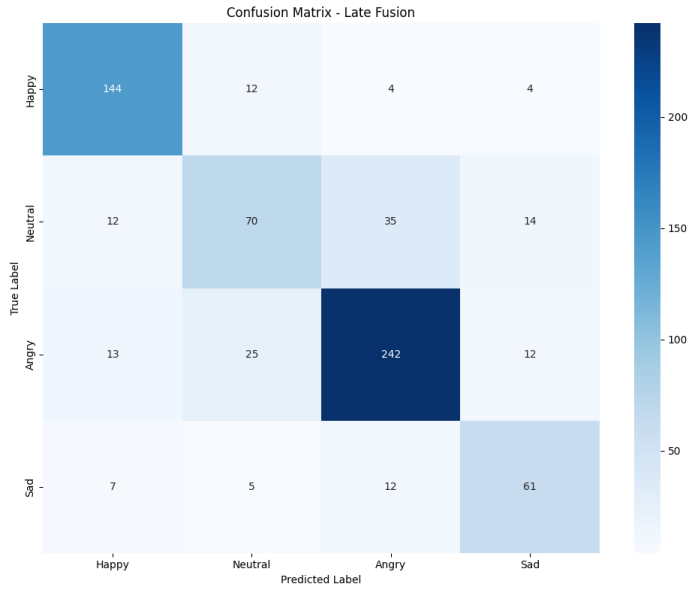

Figure 24: Confusion Matrix for the Late Fusion Model.

5.4.2 Hierarchical Attention

The Hierarchical Attention Fusion model uses attention mechanisms to process text and audio features indivitdually and combines them for classification. The Hierarchical Attention Fusion model achieved a test F1 score of **75.57%**. Figure 25 illustrates its confusion matrix.

5.4.3 Cross-Modal Transformer

The Cross-Modal Transformer Fusion model combines audio and text embeddings using a transformer-based architecture. The Cross-Modal Transformer Fusion model achieved a test F1 score of **75.37%**. Figure 26 shows its confusion matrix.

5.4.4 Gated Multimodal

The Gated Multimodal Fusion model utilizes a gating mechanism to for text and audio features to dynamically weight contributions. he Gated Multimodal Fusion model achieved a test F1 score of **76.04%**. Figure 27 presents its confusion matrix.

Figure 28 provides the comparitive analysis of training loss, validation loss, vali-

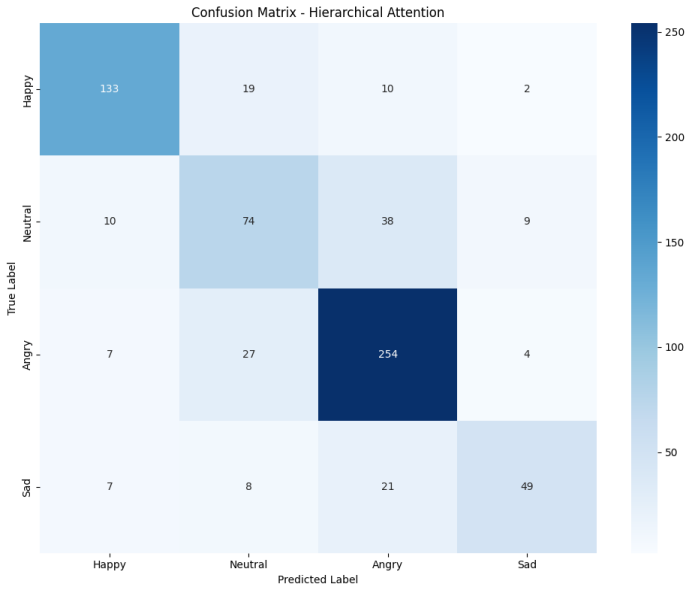

Figure 25: Confusion Matrix for the Hierarchical Attention Fusion Model.

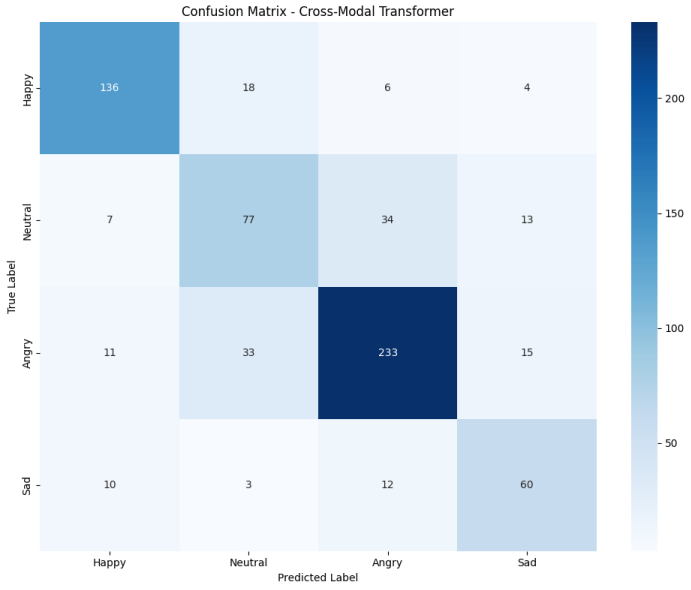

Figure 26: Confusion Matrix for the Cross-Modal Transformer Fusion Model.

dation accuracy, and validation F1 score for four models — Late Fusion, Hierarchical Attention, Cross-Modal Transformer, and Gated Multimodal across epochs.

The table 4 compares the weighted average F1 scores of the different models evaluated during the experimentation phase. Late Fusion has performed the best

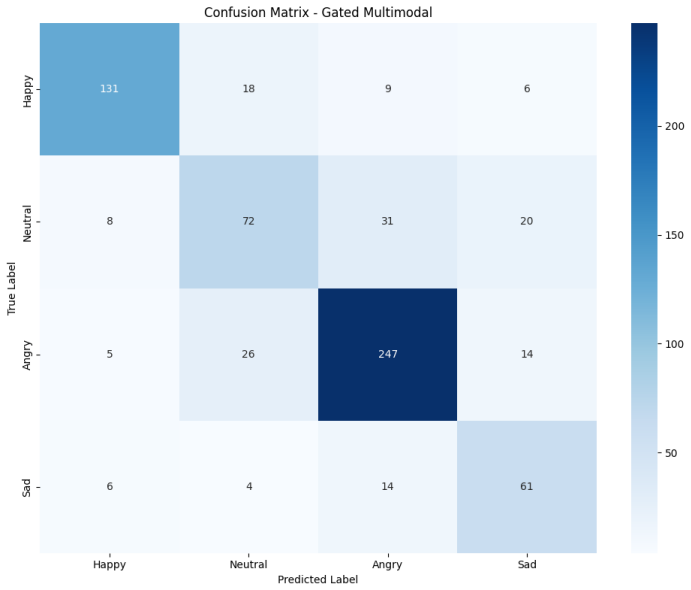

Figure 27: Confusion Matrix for the Gated Multimodal Fusion Model.

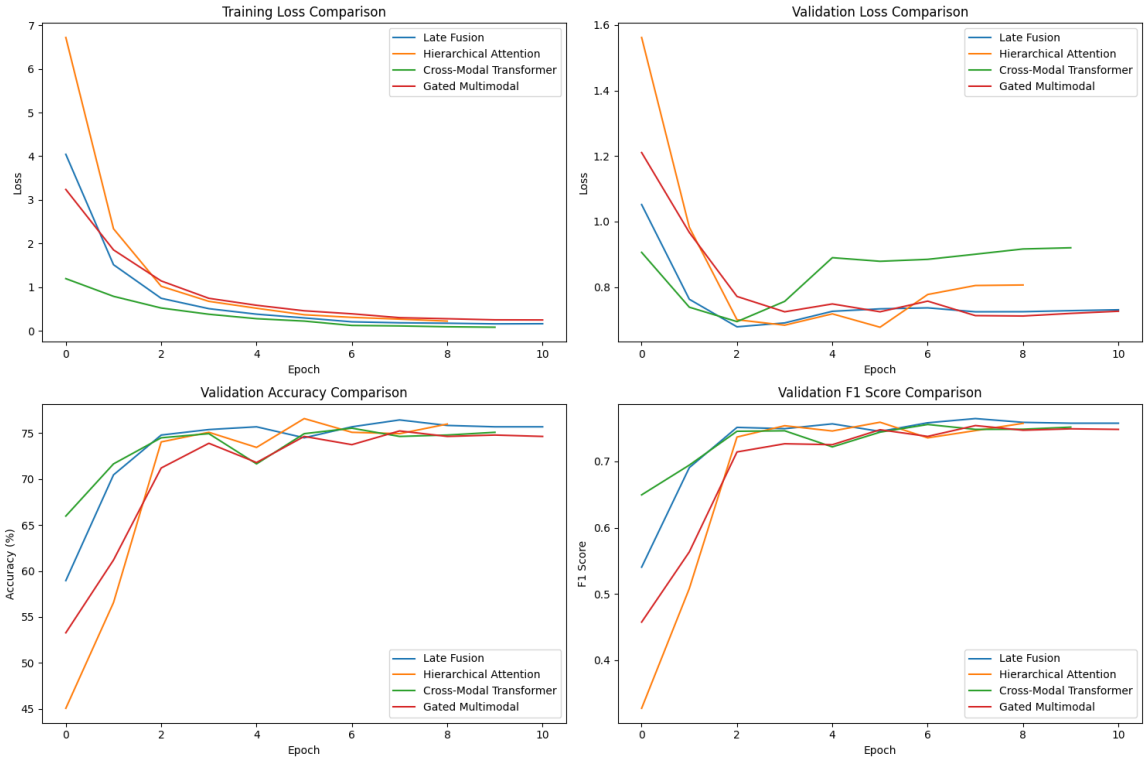

Figure 28: Comparison of training and validation performance metrics

compared to all the other models with an Weighted F1-Score of **76.67%**.

Table 3: Comparison of Different Models for IEMOCAP Dataset

ML Model	**Modal**	**Classes**	**W.F1 Score**
RoBERTa	Text	4	72.37
BERT	Text	4	69.92
RoBERTa+BiLSTM	Text	4	73.33
BERT+BiLSTM	Text	4	72.75
BiLSTM	Audio	4	62.08
AST	Audio	4	64.99
Wav2Vec2	Audio	4	66.81
Late Fusion	Text+Audio	4	76.62
Hierarchical Attention	Text+Audio	4	75.57
Cross-Modal Transformer	Text+Audio	4	75.37
Gated Multimodal	Text+Audio	4	76.04

Table 4: Comparison of Audio based Models for EMOV Dataset

ML Model	**Modal**	**Classes**	**W.F1 Score**
BiLSTM	Audio	4	97.00
AST	Audio	4	98.96
Wav2Vec2	Audio	4	99.48

CHAPTER 6

Conclusion and Future Work

This research demonstrated the potential of Fusion Models trained on multi-modal dataset when compared to other individual modals. Emotion Recognition in Conversations has been a diffitcut task and has long standing accuracy issues. Although text models were simple, fast and performed well on IEMOCAP dataset, its the Fusion models have far outperformed text based classifiers like BERT and RoBERTa with BiLSTM. Fusion approaches, which effectively combine information from multiple modalities improving overall accuracy and better handling of tasks like emotion recognition had better overall accuracy. The experimentation results indicate that while text models such as RoBERTa achieved a weighted F1 (W.F1) score of 73.37 and BERT scored 69.92, fusion models, such as the Late Fusion method and Gated Multimodal method, achieved substantially higher W.F1 scores of 76.62 and 76.04, respectively. In contrast, audio-based models underperformed on IEMOCAP and did excellent on EMOV dataset, with Wav2Vec2 reaching a W.F1 score of only 66.81 on IEMOCAP and 99.48 on EMOV datasets respectively. This highlights the complexities of conversational data like IEMOCAP.

Future prospects for this research involve expanding the study to multiple directions with additional datasets and modalities. Introducing visual modalities, such as facial expressions, could further enhance performance. Furthermore, consideration of more emotion classes, working with emotion dimensional avlues, fine-tuning existing fusion models, experimenting with advanced architectures like large-scale pre-trained multi-modal transformers, and incorporating domain-specific knowledge could further improvement emotion classification results.

Lastly, exploring real-world applications of these models, such as improving human-computer interaction in AI chatbots, healthcare systems, could demonstrate

their practical utility and impact.

LIST OF 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] P. Ekman, ''Are there basic emotions?'' *Psychological Review*, vol. 99, no. 3, pp. 550--553, 1992.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, ''Attention-based multimodal sentiment analysis and emotion detection in conversation using rnn,'' 2021.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] 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.

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