## Document Information

| | |
|---|---|
| **Analyzed document** | major_project_report_2023.pdf (D167481437) |
| **Submitted** | 5/19/2023 7:20:00 AM |
| **Submitted by** | Knowledge Resources Centre |
| **Submitter email** | krc@iiitdwd.ac.in |
| **Similarity** | 2% |
| **Analysis address** | knowledge.resources.centre.iiitdh@analysis.urkund.com |

## Sources included in the report

| | | | |
|---|---|---|---|
| W | URL: https://arxiv.org/pdf/2108.02510<br>Fetched: 9/14/2021 7:49:40 AM | ⊞ | 2 |
| W | URL: https://osf.io/preprints/inarxiv/b34gn/download<br>Fetched: 2/22/2023 8:09:06 PM | ⊞ | 6 |
| W | URL: https://dergipark.org.tr/tr/download/article-file/465485<br>Fetched: 2/1/2021 11:08:41 AM | ⊞ | 1 |
| W | URL: https://www.researchgate.net/publication/285230756_Feature_Extraction_from_Speech_Data_for_Emo...<br>Fetched: 10/31/2019 5:03:05 AM | ⊞ | 1 |

## Entire Document

i MULTIMODAL EMOTION CLASSIFICATION (TEXT AND SPEECH) Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology BY BASAVARAJ K C (19BCS018) BIPUL GAUTAM (19BCS023) DHYAN MG (19BCS038) PRANAV KM (19BCS088) May 2023 INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD

ii CERTIFICATE It is certified that the work contained in the project report titled “MULTIMODAL EMOTION CLASSIFICATION (TEXT AND SPEECH),” by “Basavaraj K C (19BCS018 )”, “Bipul Gautam (19BCS023)”, “Dhyan M G (19BCS038)” and “Pranav K M (19BCS088)” has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree. Signature of Supervisor(s) Name(s) Department(s) (Month, Year)

iii DECLARATION We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. Basavaraj K C 19BCS018 Bipul Gautam 19BCS023 Dhyan M G 19BCS038 Pranav K M 19BCS088

iv APPROVAL SHEET This project report entitled (MULTIMODAL EMOTION CLASSIFICATION (TEXT AND SPEECH)) by (BASAVARAJ K C), (BIPUL GAUTAM), (DHYAN M G) and (PRANAV K M) is approved for the degree of Bachelor of Technology in Computer Science and Engineering. Supervisor (s) _________________________ _________________________ _________________________ Head of Department _________________________ Examiners _________________________ _________________________ _________________________ Date :____________ Place:____________

ACKNOWLEDGEMENT We would like to express our deepest gratitude and appreciation to all those who have contributed to the completion of this bachelor thesis. First and foremost, We are immensely grateful to our supervisor, Dr. Nataraj K S , for his guidance, support, and valuable insights throughout this research journey. His expertise, patience, and continuous encouragement have been instrumental in shaping this thesis and my academic growth. We would also like to extend our heartfelt thanks to the faculty members and staff of Indian Institute of Information Technology, Dharwad, for providing us with a conducive learning environment and resources necessary for conducting my research. The intellectual atmosphere and the opportunities for knowledge exchange have greatly enriched my educational experience. We are deeply indebted to our family for their unwavering love, encouragement, and constant belief in our abilities. Their endless support and sacrifices have been the driving force behind my academic pursuits. Lastly, we are grateful to all the authors, researchers, and scholars whose works we have referenced in this thesis. Their contributions have laid the foundation for our research and have been invaluable in shaping our understanding of the subject matter. Basavaraj K C 19-05-2023 Bipul Gautam Dhyan M G Pranav K M .

ABSTRACT Emotions play a crucial role in human experience, and their detection and understanding can be achieved through various modalities, including facial expressions, speech, text, and body movements. In recent years, there has been a growing interest in enhancing emotion recognition accuracy by combining multiple modalities. This report focuses on the merging of speech and text features using deep neural networks to improve the accuracy of multimodal emotion classification. The Interactive Emotional Motion Capture (IEMOCAP) dataset, a comprehensive collection of audio-visual and text data, is utilized for the experimentation. This dataset contains a diverse range of emotions expressed by 10 actors in both improvised and scripted scenarios. The methodology involves analyzing speech by considering non-silent parts and extracting acoustic features from frames. Text features are generated using tokenization and word embedding techniques. Deep neural network models, such as CNN, LSTM, and LSTM with attention, are employed for classification. Feature level fusion is employed to combine features from different modalities, and separate models are trained for text and speech. A combined model is also proposed, which takes both text and speech inputs and produces the classification output. Experimental results indicate that the LSTM model achieves the highest accuracy of 69% for text classification, while the LSTM model with speech segmentation achieves the best accuracy of 47% for speech classification. However, by merging the text and speech features using a CNN model for text and a dense network model for speech, the accuracy improves to 72% for multimodal emotion recognition. The report discusses the limitations of the proposed method, including its evaluation on the IEMOCAP dataset, the need for further feature engineering, and the importance of hyperparameter tuning. It concludes by highlighting the significance of multimodal features in enhancing emotion recognition accuracy and suggests future directions for research. .

TABLE OF CONTENTS

1 CHAPTER ONE INTRODUCTION 1.1 BACKGROUND INFORMATION Emotions, as a fundamental component of the human experience, can be identified and understood through various means, including facial expressions, speech, text, and body movements. Multimodal emotion classification is a field dedicated to recognizing and categorizing emotions by utilizing multiple modalities such as text, speech, facial expressions, and physiological signals. Numerous studies and advancements have been made in this domain, with a common practice of combining different modalities to enhance the accuracy of emotion recognition. By incorporating audio and visual cues, or even integrating audio with textual information, a more comprehensive understanding of emotions can be achieved. Schuller et al. conducted a study titled "Multimodal Emotion Recognition in Speech-Based Interaction" (2011), where they explored the integration of acoustic, linguistic, and visual features for emotion recognition in speech-based interactions. Poria et al.'s research, "Multimodal Sentiment Analysis in Online Videos" (2014), focused on sentiment analysis, closely related to emotion classification. They investigated the combination of textual, visual, and acoustic modalities to analyze sentiment in online videos. Zadeh et al.'s work, "Deep Multimodal Fusion for Emotion Recognition in Intelligent Agents" (2017), proposed a deep multimodal fusion approach for emotion recognition in intelligent agents. Their study incorporated textual, acoustic, and visual modalities using deep learning techniques to enhance emotion classification performance. Caridakis et al.'s research, "Fusing Audio, Visual and Textual Cues for Emotion Recognition in User-Generated Videos" (2019), centered on emotion recognition in user-generated videos. They explored the fusion of audio, visual, and textual cues to improve emotion classification accuracy. In Zeng et al.'s survey titled "Deep Learning for Multimodal Emotion Recognition: A Survey" (2019), a comprehensive overview of deep learning techniques applied to multimodal emotion recognition was provided. The survey discussed various modalities, fusion strategies, and 2 deep learning architectures utilized in the field. It highlighted the progress made in multimodal emotion classification and identified potential areas for future research. These studies represent a subset of the extensive research carried out in the realm of multimodal emotion classification. They underscore the importance of integrating multiple modalities and harnessing advanced machine learning techniques to enhance the accuracy of emotion recognition. 1.2 STATEMENT OF THE PROBLEM The problem addressed in this research work is how to leverage deep neural networks to improve the accuracy of emotion recognition by combining speech and text features.

**63%** **MATCHING BLOCK 1/10** W

Emotion recognition is a complex task that plays a crucial role in various applications, including human-computer interaction,

virtual assistants, and sentiment analysis. However, accurately detecting and categorizing emotions solely based on individual modalities, such as speech or text, poses challenges due to the inherent limitations and variability of each modality. To overcome these limitations and enhance the accuracy of emotion recognition, there is a need to explore the potential of deep neural networks and investigate how they can effectively merge speech and text features. Our research aims to address the following questions: How can deep neural networks be designed to effectively integrate speech and text features? What architectures and techniques can be employed to optimize the fusion of these modalities? How does the integration of speech and text features using deep neural networks impact the accuracy of emotion recognition? By addressing these questions, we seek to contribute to the development of more accurate and robust emotion recognition systems that can better capture the nuances and complexities of human emotions. 3 1.3 AIM AND OBJECTIVES OF THE STUDY The aim of this research work is to develop a multi-modal emotion recognition system based on the analysis and synthesis of validated models. The study aims to address the following objectives: 1. Review and Analyze Validated Models for Emotion Detection Systems: ● Conduct a comprehensive review of existing validated models and techniques used in emotion detection systems. ● Analyze the strengths, limitations, and performance of these models in capturing emotions from various modalities, including speech and text. 2. Develop a Multi-Modal Emotion Detection System: ● Design and develop a multi-modal emotion detection system that integrates both speech and text features. ● Explore and select appropriate feature extraction methods for

**73%** **MATCHING BLOCK 2/10** W

speech region of utterance by utilizing silence removal techniques based on threshold and minimum duration of silence. ●

Investigate and implement the combination of speech features obtained through silence removal with word embeddings extracted from speech transcriptions. ● Implement appropriate algorithms and architectures, such as deep neural networks, to enable the fusion of speech and text modalities for improved emotion recognition. 3. Develop a Prototype and Evaluate Emotion Perception on IEMOCAP Dataset: ● Develop a prototype system that demonstrates and assesses emotion perception using the IEMOCAP dataset, a widely used dataset for emotion recognition research. 1.4 METHODOLOGY Following is the approach used for Multi-Modal Emotion Recognition:

4 1. Emotion Analysis in Text: Utilize models and algorithms to examine and categorize emotions in textual data. 2. Pre-processing Step: Eliminate silent segments from speech data to facilitate subsequent analysis. 3. Emotion Classification in Speech: Extract acoustic characteristics and employ machine learning algorithms to identify emotions in speech. 4. Feature Integration: Combine text and speech features to improve the accuracy of emotion recognition. 5. Assess system performance using evaluation metrics and datasets. 1.5 SIGNIFICANCE OF THE STUDY The research work on multi-modal emotion recognition holds significant importance in several aspects. The study's findings and contributions can have both theoretical and practical implications, addressing the following key areas: 1. Advancement of Emotion Recognition Technology: The study aims to contribute to the advancement of emotion recognition technology by exploring the integration of multiple modalities, specifically speech and text, in emotion detection systems. By developing and evaluating a multi-modal emotion recognition system, the study can enhance the accuracy and effectiveness of recognizing and understanding emotions from diverse sources. 2. Improved Understanding of Human-Computer Interaction: Emotion recognition plays a crucial role in human-computer interaction, particularly in applications such as virtual assistants, social robots, and affective computing. The research work can provide insights into how multimodal approaches can enhance the interaction and communication between humans and machines by enabling more nuanced and accurate emotion perception. 3. Real-world Applications and User Experience: Multi-modal emotion recognition systems have practical applications in various domains, including

5 healthcare, customer service, education, and entertainment. The study's findings can contribute to the development of emotion-aware technologies that can improve user experiences, personalized interactions, and emotional well-being in these domains. 4. Enhanced Speech Emotion Recognition: By combining speech features extracted from silence regions of utterance with word embeddings from speech transcriptions, the study aims to improve the recognition rate of speech emotion recognition systems. The research work can provide valuable insights into feature extraction techniques and fusion approaches for speech-based emotion recognition, leading to advancements in this specific domain. 5. Contribution to the Research Community: The research work can contribute to the existing body of knowledge in multi-modal emotion recognition by analyzing validated models, developing a prototype system, and evaluating its performance on the IEMOCAP dataset. The findings, methodologies, and insights gained from the study can serve as a valuable resource for researchers, practitioners, and academicians working in the field of emotion recognition and related areas. 1.6 LIMITATION OF THE STUDY While the research work on Multi Modal Emotion Recognition has its merits, it is important to acknowledge its limitations at the moment, which include: 1. Exclusion of Facial Expressions and EEG Signals: ● The study focuses on text and acoustic signals as the primary modalities for emotion recognition. ● However, it does not consider other important modalities such as facial expressions in videos and EEG signals, which can provide valuable information for a more comprehensive understanding of emotions. ● The exclusion of these modalities limits the scope and completeness of the study's findings.

6 2. Generalization to Other Datasets: ● The implementation and evaluation of the proposed approach heavily rely on the IEMOCAP dataset. ● It is crucial to recognize that the model's performance and generalizability may vary when applied to different datasets. ● The limitations of dataset diversity may restrict the extent to which the findings can be applied to real-world scenarios beyond the specific dataset used in the study. 3. Lack of Comprehensive Hyperparameter Tuning: ● Additionally, hyperparameter tuning, which plays a crucial role in optimizing model performance, is not extensively addressed. ● The absence of a thorough hyperparameter tuning may limit the study's ability to identify the most effective approaches and achieve optimal performance. It is important to acknowledge these limitations as they provide opportunities for future research to address the gaps and extend the findings of this study.

7 2 CHAPTER TWO LITERATURE REVIEW 2.1 INTRODUCTION This section offers a detailed elucidation and clarification of the concepts and terminology employed. It subsequently presents a comprehensive examination and analysis of previous research works carried out in the domain of emotion recognition, guaranteeing a comprehensive explanation of the topic. 2.2 DEFINITIONS 2.2.1 Machine Learning Machine Learning (ML) is a branch of artificial intelligence that concentrates on creating algorithms and models enabling computers to learn and make predictions or decisions autonomously, without explicit programming. ML algorithms acquire patterns and connections from data, enabling them to generalize and provide accurate predictions for new, unseen instances. An influential research paper in the ML field is "A Few Useful Things to Know About Machine Learning" by Domingos (2012). This paper offers a comprehensive overview of essential concepts and practical considerations in ML. It explores critical aspects such as the tradeoff between bias and variance, overfitting, model complexity, feature engineering, and the significance of data quality. The author underscores the importance of understanding these fundamental principles to construct effective ML models and achieve optimal performance. Another notable paper is "Deep Residual Learning for Image Recognition" by He et al. (2016). This study introduces the concept of residual networks or ResNets, which are deep neural network architectures devised to address the challenges of training extremely deep networks. The authors demonstrate the superior performance of ResNets on diverse image recognition tasks, achieving state-of-the-art results. This paper

8 highlights the potential of deep learning and the significance of architectural innovations in enhancing the capabilities of ML models. These research papers constitute only a fraction of the extensive collection of work in the field of Machine Learning. They emphasize the significance of comprehending fundamental principles, architectural advancements, and innovative approaches in advancing the field. Through these contributions, researchers continue to push the boundaries of ML, empowering computers to learn from data and provide accurate predictions or decisions across a broad range of applications. 2.2.2 Approaches Used in Machine Learning El Naqa and Murphy's research paper titled "Machine Learning in Radiation Oncology: Opportunities, Pitfalls, and Recommendations" (2015) discusses various approaches used in machine learning, including supervised learning, unsupervised learning, and semi- supervised learning. The authors provide insights into the applications and challenges of these approaches within the context of radiation oncology. Supervised Learning: According to El Naqa and Murphy, supervised learning is a widely employed approach in radiation oncology. It involves training a model using labeled data, where the input features and corresponding target labels are provided. The model learns to map the input features to the target labels based on the provided examples. Supervised learning is commonly used for tasks such as tumor segmentation, treatment response prediction, and outcome prediction. Unsupervised Learning: The research paper highlights unsupervised learning as another important approach in machine learning. In unsupervised learning, the model learns patterns and structures from unlabeled data without any explicit target labels. El Naqa and Murphy discuss how unsupervised learning techniques like clustering and dimensionality reduction can be utilized to discover hidden patterns in medical imaging data, identify subgroups of patients, and uncover relationships between variables. Semi-Supervised Learning: El Naqa and Murphy also mention semi-supervised learning as a valuable approach that combines elements of supervised and unsupervised learning. In
9 this approach, the model is trained on a combination of labeled and unlabeled data. The labeled data guides the learning process, while the unlabeled data helps to improve the model's generalization and capture more complex patterns. Semi-supervised learning is particularly useful in scenarios where acquiring labeled data is expensive or time- consuming. Figure 2.1 Machine learning algorithms that rely on input data Source: (El Naqa & Murphy, 2015) 2.2.3 Deep Learning Deep learning is a specialized area within machine learning that focuses on training artificial neural networks with multiple layers to acquire hierarchical representations of data. It has garnered significant attention and achieved notable success across diverse domains such as computer vision, natural language processing, and speech recognition. In the paper titled "Vision Transformers" by Dosovitskiy et al. (2020), a new architecture known as Vision Transformers is introduced. This approach applies self-attention mechanisms, commonly used in transformer models for natural language processing, to image recognition tasks. By leveraging self-attention, Vision Transformers demonstrate competitive performance compared to convolutional neural networks (CNNs) on various
10 image classification benchmarks. This research highlights the potency of deep learning in utilizing self-attention mechanisms to capture meaningful representations from image data. Another significant contribution is presented in the research paper "GPT-3: Language Models are Few-Shot Learners" by Brown et al. (2020). This paper unveils GPT-3 (Generative Pre-trained Transformer 3), an advanced language model based on deep learning. GPT-3 is an expansive neural network with billions of parameters, pretrained on an extensive corpus of text data. The paper showcases the exceptional capabilities of GPT- 3 in tasks related to natural language understanding and generation. Notably, GPT-3 performs well across various language-related tasks, even with minimal task-specific training data. This underscores the generalization ability of deep learning models and their capacity to learn intricate patterns from extensive pre-training. Figure 2.2 Vision Transformer Architecture Source: (Dosovitskiy et al, 2020) 2.2.4 Natural Language Processing Natural Language Processing (NLP) is an area within computer science that focuses on the interaction between computers and human language. Its objective is to develop algorithms and models that enable computers to comprehend, interpret, and generate natural language in a meaningful manner. NLP has diverse applications, including machine translation, sentiment analysis, question answering, and information retrieval.

11 One notable research paper in the field of NLP is "Natural Language Processing (Almost) from Scratch" by Collobert et al. (2011). This study introduces Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as deep learning frameworks for NLP tasks. The authors demonstrate the effectiveness of these models across various NLP tasks, such as part-of-speech tagging, named entity recognition, and semantic role labeling. The research highlights the power of deep learning methods in capturing intricate linguistic patterns and enhancing the performance of NLP systems. Another influential paper is "Distributed Representations of Words and Phrases and their Compositionality" by Mikolov et al. (2013). This research presents the Word2Vec model, which learns distributed representations, or word embeddings, from extensive text corpora. Word2Vec represents words as continuous vectors in a high-dimensional space, capturing semantic and syntactic relationships between them. The model has significantly transformed NLP by providing an effective means to encode linguistic information and improve the performance of various downstream NLP tasks, such as language modeling, sentiment analysis, and document classification. Additionally, "Attention Is All You Need" by Vaswani et al. (2017) introduces the Transformer model, which has had a profound impact on NLP. The Transformer model utilizes self-attention mechanisms to capture word dependencies within a sentence, enabling parallel processing and effective handling of long-range dependencies. The paper showcases the Transformer's superior performance in machine translation tasks and its ability to capture contextual information in a more efficient and accurate manner compared to traditional recurrent or convolutional architectures. These research papers represent only a fraction of the extensive body of work in NLP. They exemplify the advancements and breakthroughs in the field, highlighting the development of robust models and techniques that have significantly enhanced computers' understanding and generation of human language.

12 2.2.5 Speech Features Speech features play a crucial role in analyzing and understanding speech signals. These features capture different aspects of the speech signal, ranging from its temporal characteristics to spectral properties. 1. Time domain features: ● Zero Crossing Rate (ZCR): ZCR measures the rate at which the speech waveform crosses the zero axis. It provides information about the frequency content and voicing characteristics of the signal. A research paper by Emadi et al. (2016), titled "A Comparative Study of Speech Feature Extraction Techniques for Emotion Recognition," explores the use of ZCR as one of the features for emotion recognition in speech. ● Energy: Energy represents the magnitude of the speech signal at a given time. It is a measure of the overall strength or loudness of the signal. A research paper by

82% **MATCHING BLOCK 3/10** W

Reynolds and Rose (1995), titled "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models,"

discusses the use of energy as a feature for speaker identification. ● Entropy of Energy: The entropy of energy quantifies the variability or unpredictability of the energy distribution over time. It provides information about the dynamics of the speech signal. In the paper "Automatic Speaker Recognition Using Cepstral Coefficients and Hidden Markov Models" by Reynolds and Rose (1995), entropy of energy is mentioned as a feature used for speaker recognition. 2. Spectral domain features: ● Spectral Centroid: The spectral centroid represents the center of mass of the power spectrum of the speech signal. It provides information about the spectral balance or brightness of the signal. In the research paper "Automatic Genre Classification of Music Content: A Survey" by Tzanetakis and Cook (2002), spectral centroid is discussed as a feature used for genre classification of music. ● Spectral Spread: Spectral spread characterizes the spread or width of the power spectrum. It indicates the extent to which the spectral energy is distributed across

13 different frequencies. A paper titled "Audio Classification Using Modified Linear Predictive Coding Coefficients" by Ramírez et al. (2004) mentions spectral spread as a feature used for audio classification. ● Spectral Entropy: Spectral entropy measures the complexity or irregularity of the spectral distribution. It provides insights into the richness or diversity of the spectral content. In the paper "Music Genre Classification Using Wavelet Packets" by Pampalk et al. (2002), spectral entropy is discussed as a feature for music genre classification. ● Spectral Flux: Spectral flux captures the changes in the spectral magnitude over time. It represents the rate of spectral variation and is used to detect transient events or changes in the speech signal. A research paper by Dixon (2005), titled "Onset Detection Revisited," discusses spectral flux as a feature for onset detection in music signals. ● Spectral Roll-off: Spectral roll-off denotes the frequency below which a certain percentage (e.g., 90%) of the spectral energy is concentrated. It provides information about the high-frequency content or sharpness of the signal. The paper "Music Genre Classification: A Comparison of Feature Sets" by Tzanetakis and Cook (1999) mentions spectral roll-off as a feature used for music genre classification. 3. MFCCs (Mel-frequency Cepstral Coefficients): MFCCs are widely used speech features that capture the spectral characteristics of the speech signal. They are derived by applying a series of transformations, including mel-filtering and discrete cosine transformation, to the magnitude spectrum. A seminal paper by Davis and Mermelstein (1980), titled "

100% **MATCHING BLOCK 4/10** W

Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," introduces and discusses the effectiveness of MFCCs for speech recognition tasks. 4. Chroma Features: Chroma features represent the distribution of pitch classes or musical notes in the audio signal. They provide a compact representation of the harmonic content and tonal characteristics of the speech or music. The research paper "Chroma Binary Kernel

14 and its Application in Semi-Supervised Audio Classification" by Müller and Ewert (2011) explores the use of chroma features for audio classification tasks. 2.2.6 Multi-Task Learning Multi-task learning is a machine learning strategy that seeks to enhance the performance of multiple interconnected tasks by training a single model on all tasks concurrently. Its objective is to capitalize on the shared information and dependencies among tasks, thereby improving the overall learning process. An informative research article titled "An Overview of Multi-Task Learning in Deep Neural Networks" by Ruder et al. (2017) presents a comprehensive exploration of multi-task learning within the realm of deep neural networks. The paper examines various techniques, architectures, and regularization strategies employed in multi-task learning, emphasizing its advantages in terms of generalization and transfer learning. Another relevant study, "Deep Multi-Task Learning for Facial Action Unit Detection" by Liu et al. (2017), focuses on facial action unit detection, a task involving the recognition of specific facial muscle movements associated with different facial expressions. The authors propose a deep multi- task learning framework that simultaneously detects multiple facial action units, leveraging the inherent dependencies among them. The findings demonstrate enhanced performance compared to single-task learning approaches.Furthermore, Kendall et al.'s (2018) "Multi- Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics" delves into multi-task learning within the domain of scene understanding. The paper introduces a novel approach that incorporates uncertainty estimation to dynamically assign weights to the losses of different tasks during training. This adaptive weighting mechanism enables the model to allocate more attention to tasks with higher uncertainty, resulting in improved performance in both geometry and semantics tasks. These articles shed light on the potential of multi-task learning in diverse domains and provide valuable insights into its techniques and advantages. Through the simultaneous learning of multiple related tasks, multi-task learning facilitates the exploitation of shared knowledge, leading to more robust and efficient learning systems.

15 2.2.7 Multi-Modal Emotion Recognition Emotion arises from the conscious and unconscious perception of objects and circumstances, encompassing various variables like mood, behavior, personality, disposition, and motivation (Soleymani et al., 2012). Emotion plays a significant role in shaping human mental activity and thought patterns. Notably, the expression of emotion through facial, verbal, and bodily cues closely resembles the mechanism of emotional arousal. This similarity has prompted researchers to explore the possibility of assessing a subject's emotion based on its outward manifestations (Konar et al., 2015). Identifying emotions can be seen as a pattern recognition problem, as it involves recognizing the patterns associated with the expression of emotions. For instance, in speech-based emotion recognition, extracting speech features from an individual's utterances is crucial. The classification of these features into different emotion categories is known as emotion recognition. Typically, a supervised classifier takes emotional characteristics as input and predicts the corresponding emotion class as output. The human mind's mental state can be conveyed through various channels, including facial expressions, speech, gestures, posture, and biopotential signals (Konar et al., 2015). In a unimodal approach, only one mode of representation is used to recognize the emotional state, which may not always be sufficiently expressive. Relying solely on a less expressive mode can lead to misclassification. To overcome this limitation, employing multiple modalities to understand emotions can be beneficial. This approach is known as multi-modal emotion recognition (Konar et al., 2015). 2.2.8 Fusion Approaches Various fusion techniques exist for multi-modal fusion, offering different approaches at various levels of abstraction. Fusion modalities can be implemented using different methods, including low-level fusion at the signal level or intermediate-level fusion techniques such as late fusion, decision level fusion, or contextual fusion.

16 2.3 Review Of Existing Papers on Multi-Modal Emotion Recognition A. Aman Shenoy, Ashish Sardana “Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation.” Understanding emotions and sentiments in conversations is challenging because conversations are complex. The meaning of words can change based on the context in which they are spoken. Multilogue-Net acknowledges this complexity and takes it into account when analyzing emotions and sentiments. Multilogue-Net takes a multimodal approach, which means it considers different modes of communication, such as text, audio, and video. Each mode provides unique information about the conversation, and by treating them independently, the model can leverage the strengths of each modality. The fusion mechanism in Multilogue-Net combines the information from different modalities. It integrates the inputs from text, audio, and video using a fusion technique to make more accurate predictions. Multilogue-Net uses an emotion GRU (Gated Recurrent Unit) to maintain consistency in emotion detection across different tasks. The emotion GRU helps the model make consistent predictions by learning to represent emotions in a task- dependent manner. Multilogue-Net is trained using standard techniques such as cross-entropy loss and regularization. The model is evaluated on benchmark datasets, specifically the CMU Multi-modal Opinion-level Sentiment Intensity (CMU-MOSI) dataset. These datasets consist of conversations with annotated sentiment and emotion labels.

17 Figure 3.1 Pairwise attn. (fusion mechanism) Source: (Aman Shenoy, Ashish Sardana, 2020) B.

**100%** **MATCHING BLOCK 5/10** W

Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung “Multimodal speech emotion recognition using audio and text.”

Multimodal approach improves speech emotion recognition by combining multiple modalities, such as audio and text. By utilizing both audio signals and textual information, a more comprehensive understanding of speech data can be achieved. The paper introduces the Multimodal Dual Recurrent Encoder (MDRE) model, which is designed to encode audio and text information simultaneously. MDRE utilizes dual recurrent neural networks (RNNs) to encode the sequential nature of audio and text data. MDRE encodes audio and text using RNNs. For audio signals, the MDRE model employs an Audio Recurrent Encoder (ARE) that utilizes MFCC features and prosodic features. For text information, a Text Recurrent Encoder (TRE) is used, which tokenizes and indexes the speech transcripts and passes them through a word- embedding layer. Audio features are the MFCC and prosodic features. MFCC (Mel- frequency cepstral coefficients) features are widely used in audio signal analysis and provide valuable information about the spectral characteristics of the speech. Prosodic features, including F0 frequency, voicing probability, and loudness
18 contours, capture additional information related to speech intonation and rhythm. Text features include Tokenization, indexing, and word-embedding. The speech transcripts are processed using tokenization and indexing techniques to convert them into sequences of tokens. These tokens are then passed through a word- embedding layer, which maps each token to a high-dimensional vector that captures contextual meaning between words. MDRE combines audio and text information with a feed-forward neural network: This fusion allows the model to capture the complementary aspects of both modalities and make predictions based on the combined representation. Experimental results show that the proposed MDRE model achieves higher accuracy in classifying the four emotion categories compared to other existing approaches. The MDRE model addresses a common challenge in emotion recognition systems, which is the misclassification of neutral emotions. Previous models that focused only on audio features often had difficulties distinguishing neutral emotions, but the MDRE model overcomes this issue. Figure 3.2 MDREA Model Architecture Source: (Yoon, Seunghyun, Seokhyun and Kyomin Jung 2018) C.

**90%** **MATCHING BLOCK 6/10** W

Tripathi, Samarth, and Homayoon Beigi "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning"

The paper focuses on recognizing emotions by combining information from multiple modalities. It leverages speech data, text transcripts, and motion capture data to capture emotional cues from different sources. The paper explores various
19 models tailored for each modality. For speech, they experiment with MLP, LSTM models. For text, they employ 1D convolutions and LSTM models. For motion capture data, they utilize LSTM and convolutions. The paper proposes a final combined model that integrates the selected models from each modality. This fusion of models aims to leverage the strengths of each modality and improve overall emotion recognition accuracy. The authors perform hyperparameter optimization to fine-tune the combined model. They optimize the number of LSTM neurons, fully connected layer neurons, and dropout rates for each model. This process helps improve the performance of the combined model. Despite using a smaller training dataset compared to some previous studies, the proposed approach achieves competitive performance. The results show that the model performs well in recognizing emotions even with limited training data. The proposed architecture adopts a modular approach, allowing for the replacement of individual models within the combined model. This modularity enhances flexibility, as any modality-specific model can be substituted with a better-performing model without affecting the rest of the modalities. D.

**100%** **MATCHING BLOCK 7/10** W

J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak "Deep neural networks for emotion recognition combining audio and

transcripts" The paper explores the combination of acoustic features (capturing speech characteristics) and transcripts (textual representation of speech) to enhance the accuracy of emotion recognition systems. The experiments were conducted using the IEMOCAP dataset. The paper presents two individual systems: a multi- resolution Convolutional Neural Network (MCNN) and an acoustic-based Long Short-Term Memory (LSTM) network. The MCNN analyzes transcripts and incorporates different temporal contexts using parallel convolutional layers with various kernel sizes. The acoustic LSTM processes acoustic features and captures temporal dependencies in speech data. The fusion system combines the outputs or e-vectors (e.g., softmax layer outputs) from the MCNN and acoustic LSTM models. These combined representations are then fed into a Support Vector Machine (SVM) to predict the final emotion labels.

20 The MCNN outperforms previous hand-crafted e-vector features by achieving a 4% improvement in weighted accuracy (WA) for emotion recognition. Combining the MCNN and acoustic LSTM systems in the fusion approach results in a 6% relative improvement in WA compared to using either system individually. The best results are achieved by fusing all three systems together: MCNN, acoustic LSTM, and the e-vector features. This fusion approach yields a 12% improvement in WA relative to using the MCNN system alone. The fusion system's effectiveness is confirmed in experiments using Automatic Speech Recognition (ASR)-generated transcripts, which are obtained from call center data. Figure 3.3 MCNN Architecture Source: (J Cho, Raghavendra P. 2018) E.

P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding"

The researchers propose a method that utilizes both phoneme and spectrogram features to classify emotions in speech. Phonemes represent distinct units of sound, while spectrograms provide a visual representation of the speech signal's time- frequency characteristics. By combining these two types of features, the model can leverage both phonetic information and temporal-frequency patterns for more accurate emotion classification. The study finds that models based solely on spectrogram features outperform those based solely on phoneme features. The
21 researchers' combined model, which utilizes both phoneme and spectrogram features, achieves higher average class accuracy compared to existing state-of-the- art methods. This indicates that combining complementary features leads to improved performance in speech emotion classification tasks. The model effectively leverages the discriminative power of both phonetic and spectral information to better identify and classify different emotions in speech. Although spectrogram features outperform phoneme-based features in this study, phoneme-based features still hold value in certain aspects. They are particularly useful for differentiating between voiced and unvoiced speech segments and capturing semantic context. Phonemes can provide information about the presence or absence of vocal cord vibrations, as well as the intervals between sub-words and words or phrases. The proposed method has potential applications in various domains, including conversational chatbots and speech analysis for emotion recognition. Chatbots can benefit from accurately identifying and understanding the emotions expressed in speech, enabling them to respond appropriately and enhance the conversational experience. Similarly, in speech analysis tasks, such as sentiment analysis or emotion detection in audio recordings, the proposed method can contribute to more accurate and robust results. Figure 3.4 Multi Channel CNN Model(spectrogram+phoneme) Source: (P. Yenigalla et al. 2018) 3 CHAPTER THREE SYSTEM ANALYSIS AND DESIGN 22 In this chapter, we will examine the approach employed to build the deep neural network model for identifying emotions in raw audio signals. This includes discussing the tools utilized, the dataset employed, and the system requirements necessary for the construction of the model. 3.1 DATASET The Interactive Emotional Motion Capture (IEMOCAP) is a publicly available dataset. The dataset contains audio, visual and text data from 10 actors involved in improvised and scripted scenarios that elicit a wide range of emotions. Some of the important features of the IEMOCAP dataset are: ● The dataset contains 12.5 hours of audio-visual and its transcribed text data, making it one of the largest and most comprehensive datasets for emotion recognition research. ● The dataset includes various modalities, including speech, facial expressions, head and body movements, and textual transcripts. ● The dataset also includes speaker and conversational turn annotations, which can help researchers better understand the dynamics of emotion expression. This dataset has been widely used in research on emotion recognition. 3.1.1 Data Acquisition The process of obtaining and collecting relevant data for a specific task is referred to as data acquisition. In this project, it was crucial to gather data that was suitable for classifying mood in spoken utterances. To fulfill this requirement, IEMOCAP dataset was selected. This particular database is widely utilized in the field of English spoken conversation emotion datasets due to its dimensional emotion annotations. To access the dataset, a formal request was submitted electronically to the University of Southern California. It's worth noting that the data gathering process for this dataset differs from various other datasets commonly employed for emotion identification in voice utterances.
23 3.1.2 Dataset Exploratory Data Analysis (EDA) Exploratory Data Analysis (EDA) is a crucial step in understanding and analyzing a dataset before applying any machine learning or data modeling techniques. In the case of the IEMOCAP dataset, several aspects can be explored through EDA, including the distribution of gender, emotions, sessions, and the improv script. 1. Distribution of Gender: The EDA process involves examining the distribution of gender within the IEMOCAP dataset. This analysis provides insights into the representation of male and female participants in the dataset. By calculating the frequency or percentage of each gender, we can understand the gender balance and potential biases present in the dataset. Figure 3.1 Distribution of Gender 2. Distribution of Emotions: EDA also focuses on exploring the distribution of emotions in the IEMOCAP dataset. This involves identifying the different emotional states expressed by the participants in the dataset. The emotions can include categories such as happiness, sadness, anger, surprise, and others. Analyzing the distribution of emotions helps in understanding the prevalence and diversity of emotional expressions captured in the dataset.

24 Figure 3.2 Distribution of Emotions 3. Distribution of Sessions: The IEMOCAP dataset consists of recordings from multiple sessions. EDA involves examining the distribution of sessions to gain insights into the structure and composition of the dataset. This analysis helps identify any session-specific patterns, variations, or potential biases that might exist within the dataset. It also aids in understanding the context and setting in which the recordings were made. Figure 3.3 Distribution of Sessions 4. Improv Script:
25 EDA can further explore the presence and characteristics of improv scripts in the IEMOCAP dataset. This involves identifying and analyzing segments of recordings where participants engage in improvisation or spontaneous dialogue. By examining the distribution, content, and structure of the improv script segments, we can gain insights into the nature of spontaneous communication and its role in emotional expression within the dataset. Figure 3.4 Distribution of Method 3.2 REPRESENTATION AND EMBEDDING OF FEATURES The initial phase involves converting unprocessed data like text transcriptions and speech waveforms into condensed features and embeddings. These representations will be utilized as input for the system's model. The techniques for extracting features and generating embeddings for both audio and text modalities are addressed in this section. 3.2.1 Text Embeddings Text embeddings play a crucial role in natural language processing (NLP) by converting words into numerical vectors. Various methods like Word2Vec, GloVe, BERT, and FastText are used to generate these embeddings. They encode the semantic meaning and
26 interrelationships among words. Word2Vec utilizes neural networks for learning embeddings, GloVe relies on word co-occurrence statistics, BERT generates contextualized embeddings, and FastText considers subword information. By leveraging text embeddings, NLP tasks such as sentiment analysis and machine translation can be improved, enabling the development of precise and effective NLP models. Table 3.1 A compilation of widely used text embeddings: Concurrency-based Context-specific Word-level GloVe, word2vec ELMo Sentence-level doc2vec GPT, BERT Glove Embeddings GloVe is a word embedding algorithm that obtains word representations through unsupervised learning. It creates word vectors by examining the statistical properties of the entire corpus. GloVe embeddings are based on the concept that words with similar meanings often appear in similar contexts. By analyzing co-occurrence statistics, GloVe learns word vectors that capture semantic connections. These embeddings typically contain semantic and syntactic information, making them valuable for various NLP tasks like word similarity, analogy, and text classification. GloVe embeddings are pretrained on large corpora and provide fixed-size dense vectors for each word in the vocabulary. The dimensionality of GloVe embeddings is predetermined and can be adjusted based on the training data and specific needs. 3.2.2 Speech Features To extract acoustic features from speech segments in the dataset, we follow a specific algorithm. Initially, the speech files are read as vectors, and then silence removal is performed to obtain the speech segments. The algorithm for removing silence can be summarized as follows:
27 1. Define the threshold and minimum duration parameters. 2. Scan through the samples, starting from the first sample. A. If the amplitude of the current sample is below the threshold, increment a counter (n = 1). B. Repeat step 2a and increment the counter (n = n + 1) until the amplitude of the current sample is above the threshold. 3. Check the total number of accumulated samples below the threshold. A. If the number of accumulated samples (n) is greater than or equal to the minimum duration, remove those n samples. 4. Repeat steps 2-3 until all speech segments have been processed. In the context of this speech emotion recognition (SER) task, the threshold and minimum duration values play a crucial role. After conducting experiments, a threshold of 0.001% and a minimum duration of 100 ms have been found effective. It is important to note that no normalization is applied during the silence removal process, resulting in a very small

**42%** **MATCHING BLOCK 9/10** W

threshold value due to the wide dynamic range of speech signals. Once the speech segments of each utterance are obtained, feature extraction is performed on these segments. Each speech utterance is divided into frames

using a hamming window and moved with overlap steps. A set of 34 features is computed for each frame, encompassing various aspects. These features include 3

**88%** **MATCHING BLOCK 10/10** W

time domain features (zero crossing rate, energy, and energy entropy), 5 spectral domain features (spectral centroid, spectral spread, spectral entropy, spectral flux,

and spectral roll-off), as well as 13 Mel-frequency cepstral coefficients (MFCCs) and 13 chromas. 3.3 MODEL ARCHITECTURE The techniques used in the emotion recognition task are described in this section. I begin by describing the encoder model for each of the audio and text modalities separately. Then

28 I introduce a multi-modal method that uses a multi-modal model to encode both audio and linguistic information. 3.3.1 Text Model In the text classification model, Embedding Layer converts the input text into dense word embeddings using a pre-trained word embedding matrix. These embeddings can be adjusted during training to capture specific information related to the task. LSTM Layer utilizes a Bidirectional LSTM with 128 units, this layer captures sequential information in the text by processing the input in both forward and backward directions. It generates sequences of outputs for each time step.Then, Attention Decoder applies an attention mechanism to the LSTM outputs. It assigns varying weights to different parts of the input sequence, allowing the model to selectively focus on significant words or phrases. After the attention decoder, Flatten layer converts the output into a two-dimensional representation, simplifying the subsequent processing. The model employs two dense layers. The first dense layer consists of 512 units and employs the Rectified Linear Unit (ReLU) activation function to introduce non-linearity. The second dense layer consists of four units and uses the softmax activation function to produce class probabilities. 3.3.2 Speech Model The speech model begins with a Bidirectional LSTM layer consisting of 256 units, enabling it to process input sequences in both forward and backward directions. This bidirectional processing helps the model capture temporal dependencies effectively. The input shape of the LSTM layer is specified as (100, 34), indicating that each input sequence contains 100 time steps and 34 features. Subsequently, an AttentionDecoder layer is incorporated, which applies attention mechanisms to the LSTM outputs. This attention mechanism allows the model to focus on crucial parts of the input sequence while decoding. The AttentionDecoder layer comprises 128 units. To process the output from the AttentionDecoder layer, a Flatten layer is added, which reshapes the data into a one-dimensional form. Next, a Dense layer with 512 units and ReLU activation function is included in the model architecture. This layer introduces non-linearity, enabling the model to learn intricate patterns and increasing its capacity to handle complex information. The final layer is a Dense layer with 4 units,

29 utilizing the softmax activation function. It produces probabilities for each of the four possible classes, indicating the model's classification predictions. 3.3.3 Multi-modal Model The text input is transformed using an Embedding layer, converting input sequences into dense 128-dimensional vectors. Two LSTM layers with 256 units each process the embedded text sequences, generating sequential outputs. An AttentionDecoder layer captures important information within the sequence by emphasizing relevant parts. The output of the AttentionDecoder layer is further processed by an additional LSTM layer with 256 units. To prevent overfitting, a dropout layer with a rate of 0.2 is applied. A dense layer with 256 units produces the output of the text model. The speech input is a 3D tensor of shape (100, 34), representing 100 frames with 34 features per frame. Similar to the text model, two LSTM layers with 256 units each process the speech input. An AttentionDecoder layer is employed to capture important information within the sequence. The output of the AttentionDecoder layer is further processed by another LSTM layer with 256 units. To prevent overfitting, a dropout layer with a rate of 0.2 is utilized. A dense layer with 256 units generates the output of the speech model. The outputs of the text and speech models are concatenated using the concatenate function, merging the information from both modalities. The concatenated output is passed through a dense layer with 256 units and a rectified linear unit (ReLU) activation function. The combined output is passed through a dense layer with 4 units and a softmax activation function, producing predicted probabilities for the four possible classes. The model is compiled with categorical_crossentropy as the loss function, adam optimizer, and accuracy as the evaluation metric.

30 4 CHAPTER FOUR IMPLEMENTATION This section provides a summary of the results obtained from data exploration and processing, as well as the programming languages, frameworks, and libraries utilized to achieve the objectives of the project. Additionally, it encompasses the implementation details of the multi-modal model.

31 4.1 IMPLEMENTATION TOOLS USED Here, we present a summary of the different technological tools and software utilized in the creation of this system. A). Python was selected as the programming language for implementing the model due to its high level of user-friendliness and the abundance of contemporary libraries accessible to the general public. These libraries greatly facilitate the creation of machine learning models. B). NumPy, a Python extension package, serves as a replacement for NumArray and Numeric when performing numerical computations. It offers support for multi- dimensional arrays and matrices, which are commonly utilized for representing machine learning data. In my project, I employed NumPy as a storage container for efficiently managing and manipulating my training, testing, and development data, as well as conducting various numerical operations. C). Pandas, also known as the Panel Data Analysis Toolkit, is a Python library for data analysis that enhances the capabilities of analytics modeling and data manipulation. D). Matplotlib is a versatile Python library for data visualization and creating graphical charts. It is compatible with various operating systems and can seamlessly adjust even the smallest details of a figure. One of the key strengths of Matplotlib lies in its compatibility with multiple operating systems and graphics backends. This ensures that you can confidently utilize Matplotlib irrespective of your operating system or the desired output format, as it supports a wide array of backends and output formats. E). Keras is a Python-based deep learning framework that offers a convenient and accessible platform for constructing and training neural networks. It simplifies the entire process, enabling developers to swiftly prototype and explore various architectures and models. This facilitates efficient development and deployment of deep learning algorithms. Keras supports both convolutional and recurrent networks and seamlessly integrates with well-established deep learning libraries like TensorFlow and Theano. Renowned for its user-friendly nature, adaptability, and

32 scalability, Keras is widely favored by both newcomers and seasoned researchers in the field. F). TensorFlow is a machine learning framework created by Google that is available as an open-source platform. It is widely utilized for constructing and training various types of neural networks, including those used in deep learning. TensorFlow offers a flexible and efficient environment for performing numerical computations, specifically designed to handle large-scale datasets and complex mathematical operations. It provides a high-level API that facilitates the construction and training of models, while also offering a low-level API for advanced customization and fine- tuning. One of the notable features of TensorFlow is its support for both CPU and GPU computations, allowing for accelerated training and inference on compatible hardware. Due to its extensive ecosystem and strong community support, TensorFlow has gained significant popularity among researchers and practitioners in the field of machine learning. G). Pickle is a Python module that provides a simple way to serialize and deserialize objects. It allows Python objects to be converted into a binary format, making them suitable for storage in files or transmission over a network. Common use cases for pickle include saving and loading data structures, object caching, and exchanging objects between different Python processes. H). The dataset provided audio recordings in the .wav format. To handle these files in Python, the Wave module was utilized for seamless reading and writing of .wav files. I). The audiosegment Python package is a versatile tool for handling audio files of various formats. It offers an efficient and convenient way to load, manipulate, and analyze audio segments within Python code. With audiosegment, you can easily read audio files and extract specific sections or segments of interest. It supports a wide range of audio formats, such as WAV, MP3, and FLAC, making it suitable for different applications. Once the audio is loaded, audiosegment allows you to perform various operations on the segments, including applying filters, adjusting volume, converting formats, and extracting information like duration and sample

33 rate. It also enables concatenation and splitting of audio segments, facilitating work with longer files or specific tasks on smaller segments. J). Anaconda Navigator is a user-friendly graphical interface that comes bundled with the Anaconda distribution, which is a popular platform for data science and machine learning. One of the key components of Anaconda Navigator is Jupyter Notebook, a web-based interactive computing environment that allows users to create and share documents containing live code, visualizations, and explanatory text. K). Jupyter Notebook provides a flexible and intuitive interface for writing code in multiple programming languages, including Python, R, and Julia. It allows users to run code in cells, making it easy to iterate and experiment with different code blocks independently. Jupyter Notebook also supports the creation of rich content, such as charts, plots, and interactive widgets, enabling users to visualize and explore their data directly within the notebook interface. 4.2 DATA PREPROCESSING AND EXTRACTION We iterate over sessions of the dataset and retrieve the necessary paths for audio, emotion, and transcription files. We then proceed to extract the audio samples using the 'get_audio' function, transcriptions using the 'get_transcriptions' function, and emotions using the 'get_emotions' function. Figure 4.1 Extract audio samples

34 Figure 4.2 Extract transcriptions

35 Figure 4.3 Extract Emotions For each session and file, the audio samples are split into left and right channels, and the start and end times for each emotion segment are used to extract the corresponding audio samples. The extracted audio samples are associated with their respective emotion labels and transcriptions, and the resulting information is stored in the 'data' variable.

36 Figure 4.4 Splitting audio samples into left and right channels

37 Figure 4.5 Extraction of Data

38 The 'data' variable represents a structured dataset where each entry contains information about a specific emotion segment, including the audio samples, emotion label, transcription, and other relevant details. The code ensures that each entry in the 'data' variable is unique based on the ID of the emotion segment. Finally, the 'data' variable is returned as the result, containing all the collected information from the IEMOCAP dataset for further analysis or modeling tasks. It is dumped in a pickle file to be used later. 4.3 SPEECH FEATURE EXTRACTION Here, we calculate features from the audio segments in the 'data2' variable, which represents the collected information from the IEMOCAP dataset. It first sets the parameters for feature extraction, such as the window size and the overlap ratio. Then, it iterates over the audio segments, performs silence removal using a threshold-based approach, and calls the 'calculate_features' function to extract the features from the remaining voiced regions. Figure 4.6 Silence removal and features extraction The 'calculate_features' function takes the frames of the audio segment, the frequency, and additional options as inputs. It computes the short-term Fourier transform features using the 'stFeatureExtraction' function and returns the derivative of these features. If the computed features have more than two columns, the function extracts the derivative features and returns them. If the number of columns is exactly two, it returns the first column of the features. Otherwise, it returns the first column of the features as the derivative features.

39 Figure 4.7 Feature extraction The 'pad_sequence_into_array' function pads/truncates sequences to have the same length along the first dimension. Figure 4.8 Padding Sequences The resulting voiced features are stored in the 'voiced_feat' list, where each entry corresponds to the features of an audio segment. Then we save the 'voiced_feat' array as a numpy file.

40 4.4 MODEL IMPLEMENTATION A maximum sequence length is set and a Tokenizer object is initialized. The Tokenizer is then fitted on the given data. Next, the data is converted into sequences of tokens and the sequences are padded to have a fixed length of 500 using the 'pad_sequences' function from the 'sequence' module. Figure 4.9 Tokenizing Sequences Then pre-trained word embeddings from a file are loaded and stored in a dictionary called 'gembeddings_index'. A word embedding matrix, 'g_word_embedding_matrix', is created with rows representing words in the tokenizer's word index. The matrix is filled with the pre-trained word embeddings for the corresponding words.

41 Figure 4.10 Creating Glove Embedding Matrix Next, we construct the target labels by extracting the 'emotion' field from each item in the 'data2' list. One-hot encoding is applied to the target labels using the 'label_binarize' function. Figure 4.11 One Hot Encoding Labels

42 Keras, a widely used deep learning library, was utilized to build the speech, text, and concatenation models. The model architecture was constructed by incorporating layers from Keras. Figure 4.12 Implementation of deep learning text model Figure 4.13 Implementation of deep learning speech model

43 Figure 4.14 Implementation of deep learning concatenated model Figure 4.15 Text model summary

44 Figure 4.16 Speech model summary

45 Figure 4.17 Concatenated model summary 4.5 MODEL TRAINING Since the dataset lacks predefined divisions for training, development, and testing, we partitioned it into an 80:20 ratio to create separate sets for training and testing purposes. Figure 4.18 Model training

46 4.6 MODEL EVALUATION To evaluate the models, we loaded the saved models and for the corresponding input to the model(text, speech or concatenated), data preprocessing, extraction and feature extraction was done and the label was predicted by considering the highest class probability. Figure 4.19 Text Model Evaluation

47

48 Figure 4.20 Speech Model Evaluation

49

50 Figure 4.21 Concatenated Model Evaluation 4.7 MODEL RESULT We experimented with different model architectures for text, speech and combined multimodal emotions. Following are the results: Table 4.1 Unimodal Text Features Accuracy Result of the Validation Set Model Accuracy CNN 66.81% LSTM 68.6% LSTM + Attention 68% Table 4.2 Unimodal Speech Features Accuracy Result of the Validation Set Model Accuracy LSTM(Raw Data). 45% LSTM(Speech Seg.) 47% BiLSTM + Attention 46%

51 Table 4.3 Multimodal (Text+Speech) Features Accuracy Result of The Validation Set Text Model Speech Model Accuracy CNN Dense 71.5% LSTM Dense 70% LSTM LSTM 68% LSTM + Attention LSTM + Attention 68%

52 5 CHAPTER FIVE SUMMARY, FUTURE WORKS AND CONCLUSION 5.1 SUMMARY To summarize, a multi-modal method for emotion recognition is employed in this project. For text, the following steps were followed: tokenization of sentences into words, generating a 300-dimensional vector for each token, collecting vectors for each sentence, and combining text features for all sentences. Classifiers such as CNN, LSTM, and LSTM with attention were used for classification. Regarding speech, acoustic speech data was used along with the raw data. A silence removal algorithm was applied, which involved setting a threshold and minimum duration, scanning through the samples, and removing samples below the threshold based on the duration criteria. The speech data was then divided into frames and feature extraction was performed for each frame, resulting in 34 features. Classifiers such as LSTM and BLSTM with attention were used for speech classification. Feature level fusion was applied, combining features from both modalities before performing classification. Different networks were tested for each modality, and the results in terms of accuracy were observed. The models section provides details about the architecture of the text and speech classification models. The text classification model includes an embedding layer, LSTM layer, attention decoder layer, flatten layer, dense layer, and output layer. The speech classification model consists of a bidirectional LSTM layer, bidirectional attention layer, flatten layer, dense layer, and output layer. A combined model was created that takes inputs from both text and speech, and it includes layers for text processing (input, embeddings, LSTM, attention, dropout, and dense) and the same layers for speech processing. In terms of results, the LSTM model performed the best for text classification with 69% accuracy, while the LSTM with speech segmentation achieved the highest accuracy of 47%

53 for speech classification. When multimodal features were used, the accuracy of emotion recognition improved to 72% using a combination of the CNN text model and the dense speech model. Custom attention models were developed for all modalities. 5.2 FUTURE WORKS In real life, our implementation might encounter a difficulty in accurately describing emotions in a straightforward manner, which poses challenges for the annotation process and the overall recognition of emotions. The complexity and elusive nature of emotions contribute to this limitation. Additionally, environmental factors such as conversational overlap, background noise, and low-quality audio recordings significantly influence the characteristics of the input data accessible to the system. These factors have a notable impact on the overall accuracy of the model. Precisely classifying a speaker's emotional state holds the potential to greatly enhance various emotion processing systems and facilitate more natural and effective communication. Moving forward, there are several potential avenues for future research and improvement in this field: 1. Performance Evaluation: It is crucial to evaluate the performance of the proposed method on additional datasets, beyond the IEMOCAP dataset, to ensure its applicability in different contexts. For instance, exploring its performance on datasets like CMU_MOSEI can provide valuable insights. 2. Integration of Other Modalities: Further exploration can be conducted to incorporate additional modalities, such as facial expressions or physiological signals, to enhance the accuracy of emotion classification. By incorporating data from multiple modalities, the system can gain a more comprehensive understanding of emotions. 3. Exploring Alternative Architectures: Investigating alternative architectures like Transformers or Graph Neural Networks holds promise for advancing multimodal emotion
54 classification. These architectures offer new perspectives and improved performance in capturing the intricate relationships within multimodal data. 4. Impact of Pre-processing Techniques: Examining the effects of different pre- processing techniques, such as noise removal or speaker normalization, is essential to refine the system's accuracy. These techniques can help mitigate the negative impact of environmental factors and enhance the quality of the input data. 5. Utilization of BERT: BERT, a powerful pre-trained language model, presents an intriguing avenue for future research. Fine-tuning a pre-trained BERT model using speech data annotated with emotions could yield valuable insights and improve the performance of speech emotion recognition systems. 6. Modeling Emotional Shifts: Emotion recognition is further complicated by the dynamic nature of emotions, which can rapidly change over time. Future work can focus on developing methods to model emotional shifts within speech. One potential approach is integrating an "emotional shift component" into the proposed model, enabling a more nuanced understanding of emotional transitions. By addressing these areas of future research, significant progress can be made in the field of multimodal emotion classification. This progress will lead to improved accuracy in emotion recognition systems, ultimately facilitating more effective and empathetic human- computer interaction. 5.3 CONCLUSION In conclusion, our research focused on developing a multimodal emotion classification system that combines both acoustic and text features. By employing an early fusion technique, we merged the two sets of features and explored different architectures for each network: text, acoustic, and concatenation. Through our experiments, we found that the best performance was achieved using LSTM for text classification, LSTM with speech segmentation for speech classification, and a combination of CNN and dense network for the overall multimodal approach.
55 The incorporation of multimodal features resulted in a significant enhancement in accuracy for categorical emotion recognition. Specifically, we observed a remarkable increase of 24.5% in accuracy compared to using only acoustic features. Additionally, the multimodal approach provided an extra 3% improvement in accuracy compared to utilizing only text features. These findings highlight the importance of leveraging multiple modalities, such as speech and text, in capturing a more comprehensive understanding of human emotions. The combination of acoustic and textual information enables a more nuanced and accurate emotion classification system. Our research contributes to the growing body of knowledge in the field of multimodal emotion recognition and emphasizes the effectiveness of multimodal fusion techniques. Overall, our study demonstrates the significance of multimodal approaches in emotion classification, paving the way for advancements in various domains, including human- computer interaction, affective computing, and psychological research. By understanding and accurately recognizing emotions, technology can better adapt to human needs and provide more personalized and empathetic experiences.
56 6 REFERENCES

## Hit and source - focused comparison, Side by Side

| | |
|---|---|
| **Submitted text** | As student entered the text in the submitted document. |
| **Matching text** | As the text appears in the source. |

| 1/10 **SUBMITTED TEXT** 18 WORDS | **63%** **MATCHING TEXT** 18 WORDS |
|---|---|
| Emotion recognition is a complex task that plays a crucial role in various applications, including human-computer interaction, | emotion recognition (SER) is a challenging task that plays a crucial role in natural human-computer interaction. |

W https://arxiv.org/pdf/2108.02510

| 2/10 SUBMITTED TEXT 20 WORDS | 73% MATCHING TEXT 20 WORDS |
|---|---|
| speech region of utterance by utilizing silence removal techniques based on threshold and minimum duration of silence. ● | speech region of utterance which is obtained by silence removal technique based on threshold and minimum duration of silence. |

W https://osf.io/preprints/inarxiv/b34gn/download

| 3/10 SUBMITTED TEXT 15 WORDS | 82% MATCHING TEXT 15 WORDS |
|---|---|
| Reynolds and Rose (1995), titled "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," | Reynolds and R.C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", |

W https://dergipark.org.tr/tr/download/article-file/465485

| 4/10 SUBMITTED TEXT 13 WORDS | 100% MATCHING TEXT 13 WORDS |
|---|---|
| Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," | Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences |

W https://www.researchgate.net/publication/285230756_Feature_Extraction_from_Speech_Data_for_Emotio ...

| 5/10 SUBMITTED TEXT 16 WORDS | 100% MATCHING TEXT 16 WORDS |
|---|---|
| Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung "Multimodal speech emotion recognition using audio and text." | Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal Speech Emotion Recognition Using Audio and Text." |

W https://osf.io/preprints/inarxiv/b34gn/download

| 6/10 SUBMITTED TEXT 15 WORDS | 90% MATCHING TEXT 15 WORDS |
|---|---|
| Tripathi, Samarth, and Homayoon Beigi "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning" | Tripathi Sarthak, and Homayoon Beigi. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. |

W https://arxiv.org/pdf/2108.02510

| 7/10 SUBMITTED TEXT 23 WORDS | 100% MATCHING TEXT 23 WORDS |
|---|---|
| J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak "Deep neural networks for emotion recognition combining audio and | J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, Deep neural networks for emotion recognition combining audio and |

W https://osf.io/preprints/inarxiv/b34gn/download

| 8/10 SUBMITTED TEXT 22 WORDS | 100% MATCHING TEXT 22 WORDS |
|---|---|
| P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding" | P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, Speech Emotion Recognition Using Spectrogram & Phoneme Embedding, |
| W https://osf.io/preprints/inarxiv/b34gn/download | |

| 9/10 SUBMITTED TEXT 35 WORDS | 42% MATCHING TEXT 35 WORDS |
|---|---|
| threshold value due to the wide dynamic range of speech signals. Once the speech segments of each utterance are obtained, feature extraction is performed on these segments. Each speech utterance is divided into frames | threshold value is very small due to the wide dynamics of the speech signal. After getting the speech segment of each utterance, we perform feature extraction based on those speech segments. Each speech utterance is split into frames |
| W https://osf.io/preprints/inarxiv/b34gn/download | |

| 10/10 SUBMITTED TEXT 23 WORDS | 88% MATCHING TEXT 23 WORDS |
|---|---|
| time domain features (zero crossing rate, energy, and energy entropy), 5 spectral domain features (spectral centroid, spectral spread, spectral entropy, spectral flux, | time domain features (zero crossing rate, energy, and the entropy of energy), 5 spectral domain features (spectral centroid, spectral spread, spectral entropy, spectral flux, |
| W https://osf.io/preprints/inarxiv/b34gn/download | |