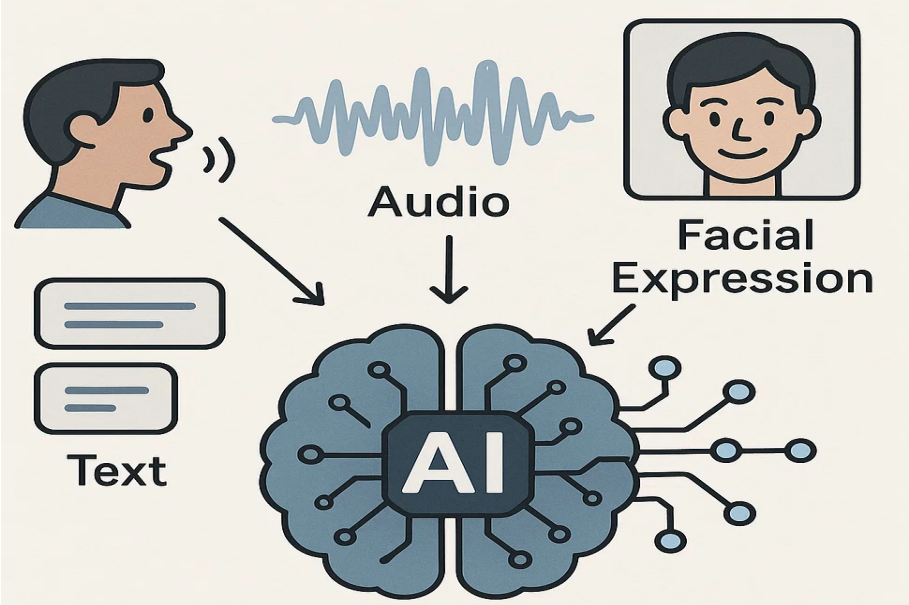
ISEL



Emotion Recognition in Multimedia Content

SOFIA FERNANDES CONDESSO
(Licenciada em Matemática Aplicada)

Dissertação para obtenção do grau de Mestre em Engenharia Informática e Multimédia

Orientadores:
Doutor Artur Jorge Ferreira
Doutor Nuno Miguel da Costa de Sousa Leite

Júri:
Presidente: Doutor Pedro Miguel Torres Mendes Jorge
Vogais:
Doutor Gonçalo Caetano Marques
Doutor Artur Jorge Ferreira

Dezembro de 2025

Emotion Recognition in Multimedia Content

SOFIA FERNANDES CONDESSO
(Licenciada em Matemática Aplicada)

Dissertação para obtenção do grau de Mestre em Engenharia Informática e Multimédia

Orientadores:
Doutor Artur Jorge Ferreira, DEI/ISEL
Doutor Nuno Miguel da Costa de Sousa Leite, DEI/ISEL

Júri:
Presidente: Doutor Pedro Miguel Torres Mendes Jorge, DEI/ISEL
Vogais:
Doutor Gonçalo Caetano Marques, DEI/ISEL
Doutor Artur Jorge Ferreira, DEI/ISEL

Dezembro de 2025

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Artur Ferreira and Professor Nuno Leite, for their continuous guidance and support throughout the development of this work. Professor Artur, who was also my lecturer in *Image Processing and Biometrics* — my first real contact with the field of image processing — and *Learning and Data Mining* has been an inspiring mentor whose attention to detail and clarity in teaching helped me approach challenges with confidence. Professor Nuno's insightful feedback and direction were also essential in shaping both the structure and the quality of this thesis.

I am also grateful to the Instituto Politécnico de Lisboa (IPL) and the Instituto Superior de Engenharia de Lisboa (ISEL) for providing the academic environment and resources that made this research possible.

This work originated from the *Lumios* project, and again, I am deeply thankful to my supervisors for embracing this idea with enthusiasm and openness. I would like to express my appreciation to my colleagues from *Lumios* — Roberto Franzan, Professor Giandomenico Iannetti, and Anna Kopach — for their valuable collaboration and inspiration. Roberto, whose idea formed the foundation of this research, provided essential motivation and technical insight; Professor Giandomenico's academic expertise and stimulating discussions enriched my perspective of this work; and Anna's thoughtful contributions on product management brought a practical and applied perspective to the project.

Finally, I would like to express my heartfelt thanks to Miguel for his constant support and understanding throughout this journey, and I am especially grateful to my mother, Manuela, for always being available to listen and reflect with me. Their patience, encouragement, and care were essential during the more demanding moments of this process.

Statement of integrity

I declare that this dissertation is the result of my personal and independent research. Its content is original, and all sources listed in the bibliographic references were consulted and are duly mentioned in the text. I further declare that all scientific and technical references relevant to the development of the work are duly cited and included in the bibliographic references.

The author

Sofia Fernandes Condesso

Lisbon, December 18 th, 2025

Abstract

Emotion Recognition (ER) has become crucial in Human-Computer Interaction (HCI), with applications ranging from mental health support to adaptive learning. While many existing approaches rely on controlled environments or hardware-based sensors, this thesis explores non-contact unimodal methods—speech, facial expressions, and textual data—for a more naturalistic and practical analysis of emotions.

First, we conduct a systematic evaluation of unimodal ER, comparing classical Machine Learning (ML) and Deep Learning (DL) approaches across multiple unimodal and multimodal datasets. For speech modality (audio), we extract acoustic features using openSMILE (GeMAPS), and learn with models such as Support Vector Machines (SVM) and Random Forests. Results show that feature selection on acoustic features can improve Speech Emotion Recognition (SER). For Facial Emotion Recognition (FER), we experiment with DeepFace and a lightweight Convolutional Neural Networks (CNN). For textual emotion recognition, we employ Word2Vec and GloVe with ML and DL models, and also experiment zero-shot and few-shot learning with large language models. In multimodal experiments, fusion of text and audio modalities improved accuracy to 0.45, confirming the benefit of combining complementary emotional cues. However, adding the visual modality led to a slight degradation in performance, attributed to suboptimal frame sampling. Overall, results highlight the trade-offs between unimodal simplicity and multimodal robustness, demonstrating that lightweight, interpretable models can achieve practical performance for real-world emotion-aware applications.

Key words

Acoustic Features; Deep Learning; Emotion Recognition; Facial Expressions; Human-Computer Interaction; Large Language Models; Machine Learning; Multimodal Data.

Resumo

O Reconhecimento de Emoções (RE) tem vindo a ganhar importância na Interação Humano-Computador (IHC), com aplicações em áreas como saúde mental, ensino adaptativo e interfaces inteligentes. Muitos métodos existentes dependem de ambientes controlados ou sensores físicos; esta dissertação explora abordagens não intrusivas e unimodais — baseadas em fala, expressões faciais e texto — para uma análise mais naturalista e acessível das emoções. Numa primeira fase, é realizada uma avaliação sistemática de métodos unimodais de RE, comparando algoritmos de *Machine Learning* (ML) e *Deep Learning* (DL) em diversos conjuntos de dados unimodais e multimodais. Para a modalidade de fala (áudio), são extraídas características acústicas com a biblioteca *openSMILE* (*GeMAPS*), e o treino executado com modelos tradicionais, como *Support Vector Machines* (SVM) e *Random Forests*. Verifica-se que a seleção de características melhora o desempenho em Reconhecimento de Emoções na Fala (REF). Para o Reconhecimento Facial de Emoções (RFE), são explorados modelos como o *DeepFace* e uma *Convolutional Neural Network* (CNN) leve. No caso do Reconhecimento de Emoções em Texto (RET), são utilizados *Word2Vec*, *GloVe* e abordagens baseadas em *zero-shot* e *few-shot learning* com *Large Language Models* (LLM). Na fusão multimodal, a combinação das modalidades de texto e áudio aumentou a taxa de acerto para 0.45, demonstrando a utilidade de combinar diferentes fontes de informação emocional. A adição da modalidade visual resultou numa ligeira degradação de desempenho, atribuída à estratégia de amostragem de frames. Os resultados revelam o equilíbrio entre a simplicidade dos métodos unimodais e a robustez das abordagens multimodais, mostrando que soluções leves e interpretáveis podem alcançar desempenho competitivo em aplicações reais de reconhecimento emocional.

Palavras chave

Aprendizagem Automática; Aprendizagem Profunda; Características Acústicas; Dados Multimodais; Expressões Faciais; Interação Humano-Computador; Large Language Models; Reconhecimento de Emoções.

Table of Contents

List of Figures

List of Tables

Acronyms

ASR Automatic Speech Recognition. xvii, 50

AU Action Unit. xvii, 35

BiLSTM Bidirectional Long Short-Term Memory. xvii, 30

BoW Bag-of-Words. xvii, 21

BPE Byte-Pair Encoding. xvii, 20, 21

CLAHE Contrast Limited Adaptive Histogram Equalization. xvii, 60

CNN Convolutional Neural Networks. xvii, 30, 31, 33–35, 37, 64, 68

CPU Central Processing Units. xvii, 47, 50, 51, 73

DBN Deep Belief Networks. xvii, 42

DL Deep Learning. xvii, 36

eGeMAPS Extended Geneva Minimalistic Acoustic Parameter Set. xvii, 26, 27, 29

ER Emotion Recognition. xvii, 6, 9, 24, 71

ERC Emotion Recognition in Conversations. xvii, 24, 25

FACS Facial Action Coding System. xvii, 35

FER Facial Emotion Recognition. xiii, xvii, 6, 9, 12, 15, 33, 34, 36, 37, 80

FP Feature Point. xvii, 35

GCN Graph Convolution Networks. xvii, 30, 31, 41

GeMAPS Geneva Minimalistic Acoustic Parameter Set. xvii, 26, 27, 29

GMM Gaussian Mixture Models. xvii, 29, 30, 42

GPU Graphics Processing Units. xvii, 73

HCI Human-Computer Interaction. xvii, 2, 3

HMM Hidden Markov Models. xvii, 29, 30, 36

HNR Harmonics-to-Noise Ratio. xvii, 28, 29

HOG Histogram of Oriented Gradients. xvii, 35

KNN K-Nearest Neighbours. xvii, 29

LBP Local Binary Pattern. xvii, 34

LDP Local Directional Pattern. xvii, 34, 35

LLD Low-Level Descriptor. xvii, 27, 29

LLM Large Language Models. xiii, xv, xvii, 2, 5, 7, 19, 22–25, 41, 43, 47, 52–54, 71, 73, 76, 79

LPCC Linear Predictive Cepstral Coefficients. xvii, 28

LSTM Long Short-Term Memory. xvii, 30, 37

MER Multimodal Emotion Recognition. xvii, 6, 9, 16, 39, 40, 42, 43

MFCC Mel-Frequency Cepstral Coefficients. xvii, 28, 31, 76

ML Machine Learning. xvii, 18

NLG Natural Language Generation. xvii, 6, 7

NLP Natural Language Processing. xvii, 19–21

NLU Natural Language Understanding. xvii, 6

OOV Out-Of-Vocabulary. xvii, 20

PCA Principal Component Analysis. xvii, 30, 42

RAG Retrieve Augmented Generation. xvii, 23

RNN Recurrent Neural Networks. xvii, 30, 37

RvNN Recursive Neural Networks. xvii

SA Sentiment Analysis. xvii

SDK Software Development Kits. xvii, 4, 5

SER Speech Emotion Recognition. xvii, 6, 9, 12–15, 22, 25–27, 29, 31, 43, 56

SIFT Scale-Invariant Feature Transform. xvii, 35

STFT Short-Time Fourier Transform. xvii

SVM Support Vector Machines. xvii, 22, 29–31, 33, 37

TER Textual Emotion Recognition. xvii, 6, 9, 12, 14, 19, 20, 24, 48

TF-IDF Term Frequency–Inverse Document Frequency. xvii, 21, 22

WER Word Error Rate. xvii, 50

Chapter 1

Introduction

The ability to recognize and understand emotions has become crucial in today's society, which demands that we know how to deal with our emotions effectively. In this context, the development of tools to help us better understand our emotions is fundamental to help people deal with emotional events and contribute to personal and professional growth. This thesis explores the development of an application designed to assist auto-assessment, while leveraging machine learning to provide users with actionable insights into their emotional communication. Instead of focusing on achieving state-of-the-art results, this thesis focuses on developing a solution to help users analyze expressed emotions from their multimedia artifacts, improving their ability to self-regulate emotional states given the occasion or scenario. Nonetheless, this thesis will review the concepts and approaches of all components of the proposed application.

1.1 Emotion Recognition in Multimedia Content

1.1.1 Definition

The definition of multimedia has evolved depending on perspective. For hardware vendors, it has often meant computers equipped with sound cards and video capabilities. For entertainment providers, multimedia refers to interactive broadcasting and on-demand services. In computer science, multimedia is generally defined as artifacts or applications that integrate multiple modalities, such as text, images, graphics, animation, audio, video, and often interactivity into a unified system [5].

Historically, multimedia communication can be traced through different mass media innovations: newspapers combining text and graphics, motion pictures, silent films, radio transmissions, and television in the twentieth century.

Recognizing emotions from multimedia content has been a topic of research since the mid

1980s, when statistical properties of certain acoustic features were used to obtain insights about the emotional state of the speaker. Only later, with the advent of more powerful computers, this field started to use visual cues to recognize emotions from image or video data. Nowadays, television and internet allow the distribution of new types of content, technologies and means of communication that are now available to everyone's home, and ultimately, to everyone's hand palm. As time passes, the definition of multimedia became broader, currently including all types of information, namely text, audio, image, and video content, which encompass a variety of emotional states and messages.

Emotion recognition in multimedia content is the process that aims to recognize the emotions present in a given multimedia input. Modality is another important concept, which is defined as the type of information and subsequent representation format in which information is stored.

1.1.2 Motivation

The detection of emotions in multimedia content is a field that has advanced significantly in recent years, making possible some innovative applications in a variety of areas, from Human-Computer Interaction (HCI), education, and entertainment, to healthcare, security and customer service, to name a few. However, the analysis and interpretation of complex emotions is still a present challenge of interest. The use of Large Language Models (LLM) to analyze information and produce textual content with certain characteristics, or the possibility of talking to a chatbot, are solutions that are being increasingly used, both by organizations and for personal use. Therefore, the product of this research can be used by individuals to improve social skills and to help with emotion regulation and public presentations.

The problem of social anxiety and depression affects more some countries than others. For example, in Japan, a bigger problem has emerged and became the subject of research, named "Hikikomori", which is characterized by individuals withdrawing from society and isolating themselves for extended periods, often retreating to their rooms for months or even years at a time [6]. These individuals could benefit from tools that allow them to assess their expressed emotions, when communicating with others.

Traditional approaches to emotion recognition often rely on physiological sensors, such as electrodermal activity monitors or heart-rate trackers, which require direct physical contact with the subject [7]. While such methods can provide precise signals of autonomic nervous system activity, they are intrusive in nature and may themselves alter the emotional state being assessed. For individuals experiencing social anxiety or related conditions, the presence of sensors or wearable devices can increase discomfort, compromising the goal of unobtrusive monitoring.

In contrast, this thesis focuses on non-contact modalities, audio, image, and text, which allow emotion analysis in a naturalistic and non-intrusive manner. This choice reduces the participant burden and makes the system more practical for everyday use.

Emotion recognition in text, audio, and image is a well-known problem, with state-of-the-art solutions that range from classical machine learning models to deep learning solutions, using smaller and larger architectures.

The combination of different modalities makes it possible to capture emotional nuances that would be unnoticeable through the analysis of only one of the modalities [8]. In this thesis, we will focus, first, on recognizing emotions on all available modalities, using lightweight solutions that performed well on established benchmarks then, we will generate a feedback report stating the recognized emotions, the time at which emotion events occurred, and some final considerations about the overall results, this may serve as a starting point for chatbot interaction that is prompted in the end of the normal usage scenario.

1.1.3 Applications

Human-Computer Interaction (HCI) and Customer Service: There are many HCI applications, for example, customer service bots adjust their responses based on the emotions present in the customer's voice [9]. With this, call centers can better adjust responses to improve the level of customer satisfaction. An example of this type of application is Deloitte's TrueVoice [10].

Customer Segmentation and Marketing: In the commercial sector, emotion recognition is being explored as a tool to optimize marketing strategies and improve customer service. By analyzing emotional responses to ads and products, marketers aim to improve targeting and engagement. Kairos [11] is a company specialized in serving businesses with face recognition, but their platform also includes emotion detection capabilities to analyze consumer interactions.

Mental Health Monitoring: Emotion recognition systems are promising as auxiliary tools in psychological and neuroscientific research. These systems can potentially help in clinical practice, namely in the diagnosis and monitoring of neuropsychiatric disorders, by providing objective and quantifiable data on patient's emotional responses which allows efficient patient screening and appropriate treatment planning [12, 13, 14]. However, it is crucial to recognize the limitations of current systems, particularly with regard to accurately capturing the complexity of human emotions. In addition, the use of automated systems requires rigorous validation and careful interpretation of results to avoid misdiagnosis or over-reliance on algorithmic results.

Education and Work: Learning software prototypes have been developed to adapt to kids' emotions. When the child shows frustration because a task is too difficult or too simple, the

program adapts the task so it becomes less or more challenging. Another example is a system that adjusts lighting and music of the office, based on the mood of the people in the room and creates an environment that motivates them to work.

Surveillance and Security: The integration of facial and emotion recognition algorithms into surveillance and security systems is a highly controversial area. Proponents argue that these technologies can increase public safety by identifying individuals of interest or predicting potentially harmful behavior, while dissenters pose that the accuracy and reliability of these systems in real-life scenarios are still the subject of important debate. The main concerns with this application are related to possible violations of privacy and the potential for discriminatory practices, so the deployment of such systems in public spaces and workplaces raises serious ethical questions that require careful analysis and framing within current regulations. On the other hand, controlled environments, such as interrogation rooms, can offer a controlled application, where technology can potentially provide information about a suspect's emotional state, respecting strict ethical and legal limits.

Commercial and Enterprise-Grade Software Solutions: Several companies have developed robust software for real-time emotion recognition. These platforms typically use proprietary algorithms and are often integrated into larger human behavior analytics or marketing research systems.

iMotions is an IT company specialized in human behavior research software and is the provider of a hardware-agnostic platform that integrates and synchronizes multiple biosensors, such as eye tracking, facial expression analysis, and physiological signals, into a single application [15]. Emotient is a facial expression analysis engine that is now integrated in the iMotions platform, providing detailed data on micro-expressions and facial action units. Affectiva [16] develops software to recognize human emotions and cognitive behavior from facial cues and voice, and has also recently been integrated into the iMotions platform.

EmoVu, developed by Eyeris [17], is a deep learning-based emotion recognition system specialized in reading facial microexpressions in real-time. Eyeris focuses on providing high-fidelity emotional data for in-cabin automotive safety and consumer research.

NVISO is an European company focused on cyber-security that also researches on emotional and behavioral detection and prediction. Their technology, which has been presented in a partnership with BrainChip [18], is a key player in the detection of emotional and behavioral cues for various applications, including robotics and interactive displays.

For researchers and developers, several open-source tools and Software Development Kits (SDK) are available. These provide the foundational code for integrating emotion recognition

into custom applications. Cognitive-Emotion-Python [19] is a good example of a Python SDK for the Microsoft Emotion API, integrated with Cognitive Services.

1.2 Problem Statement

Although the detection of emotions in multimedia content has seen significant advancements, the nuanced interpretation of complex emotional patterns remains a persistent challenge. Traditional approaches often fail in providing personalized feedback that translates into tangible improvements in communication. However, the emergence of LLM presents a unique opportunity to classify emotions and to provide contextually rich interpretations and interactive guidance. This research addresses the need for a system that goes beyond simple emotion detection, offering users a conversational partner to explore and understand their emotional communication.

1.2.1 Research Question

The central research question addressed in this thesis is "*How to develop a lightweight artificial intelligence system that, through the analysis of multimodal data (voice, video and text), provides personalized feedback on expressed emotions and communication patterns, facilitating self-reflection and personal development?*".
To address this question, our aim is to: *(i)* develop a robust and accurate multimodal emotion recognition system, *(ii)* leverage LLM to provide users with meaningful and personalized feedback, *(iii)* create an engaging user experience that promotes self-awareness, and *(iv)* provide actionable advice for improving communication. This research focuses on creating a modular system that is applicable in various applications with a rigorous evaluation of performance on established benchmarks.
We will review the existing literature which will guide us to an informed choice of techniques to use in each component of our system. This choice will be made by trading off these goals:

1. Minimize the time taken on models' inferences and fusion.

2. Maximize performance in the main benchmarks.

The focus is on creating a module that can be applied in multiple applications, presenting and discussing the results obtained from various tasks involved in developing the proposed system.

1.2.2 Task Definition

The core problem addressed in this research is the development of a system capable of accurately detecting and interpreting emotions from multimodal data (audio, image, or text) to provide users with personalized feedback and insights. This involves two main tasks: the emotion recognition task and the conversational AI task.

The task of Emotion Recognition (ER) consists of determining the emotional state of a given sample, which can be an image, a textual utterance or an audio utterance. Formally, given a sample M represented by its feature vector $x \in \mathbb{R}^F$, where F is the dimensionality of the feature space, the goal is to predict the emotion label e from a predefined set of emotion labels $C = \{c_1, c_2, ..., c_K\}$, where K is the number of distinct emotion classes. This involves learning a function f that maps the input feature vectors x to the predefined set of emotions C.

Formally, the task for which we are aiming is Multimodal Emotion Recognition (MER), where the unimodal subtasks can include Facial Emotion Recognition (FER), Speech Emotion Recognition (SER), and Textual Emotion Recognition (TER).

A conversational dataset can be defined as $C = \{(U_1, R_1, S_1), ..., (U_N, R_N, S_N)\}$, where each tuple consists of a user utterance U_j, a system response R_j, and a dialogue state S_j representing context up to turn j.

Each user utterance U_j can be represented as a sequence of feature vectors $u_j = (u_{j1}, ..., u_{jn})$, where each $u_{ji} \in \mathbb{R}^p$ represents a feature vector for the i-th word or token in U_j, and n is the length of the utterance. Similarly, each system response R_j can be represented as a sequence of feature vectors $r_j = (r_{j1}, r_{j2}, ..., r_{jm})$, where each $r_{ji} \in \mathbb{R}^q$ represents a feature vector for the i-th word or token in R_j, and m is the length of the response. The dialogue state S_j can be represented as a feature vector $s_j \in \mathbb{R}^d$, where d is the dimensionality of the dialogue state representation.

The task consists in learning a function f that maps a user utterance U_j and a dialogue state S_{j-1} (or S_0 for the first turn) to a system response R_j and an updated dialogue state S_j, given the dataset C. Formally:

$$f(U_j, S_{j-1}) \rightarrow (R_j, S_j)$$

This task involves Natural Language Understanding (NLU), to interpret the user's utterance U_j and to extract relevant information, such as intent and entities. It is also needed to maintain and update the dialogue state S_j based on the user utterance and the previous dialogue state. Lastly, it involves Natural Language Generation (NLG), providing a coherent and relevant system response R_j based on the dialogue state.

1.2.3 Proposed Solution

This work proposes the development of a system capable of detecting and analyzing emotions such as anger, joy, sadness, fear, or surprise in multimodal content (audio and text). Emotions are detected by training classification models for both modalities. The results are then analyzed by an LLM-powered "agent" through chat-bot interaction. This agent utilizes its contextual understanding and NLG capabilities to generate a report summarizing the emotional content and allows users to interact with the system, asking questions and receiving personalized feedback on their emotional expression and communication patterns. We propose to use the LLM ability to interpret complex emotional data expressed on the text that corresponds to the transcript of the uploaded content and to generate human-like responses. This approach will make it ideal for providing users with a nuanced understanding of their emotional states.

1.2.4 Ethical Concerns

The task of emotion recognition and its applications carry concerns related to the privacy and consent of both the application users and individuals related to the data to train the models. Developers and stakeholders should focus on what systems should do and set limits on their action through consent.

Although these applications offer potential benefits in terms of efficiency and personalization, it is crucial to consider the ethical implications of using emotion recognition for commercial purposes, particularly in terms of data privacy and potential manipulation.

It is also important that the accuracy of the developed systems is as reliable as possible for all demographic groups, since the purpose of such systems is to facilitate self-knowledge and emotion regulation, it is very undesirable for the user to get inaccurate feedback and to possibly feel discriminated.

The EU AI Act [20] (2024) defines biometric identification as "*automated recognition of physical, physiological and behavioural human feature such as the face, eye movement, body shape, voice, prosody, gait, posture, heart rate, blood pressure, odor, keystrokes, characteristics, for the purpose of establishing an individual's identity by comparing biometric data of that individual to stored biometric data of individuals in a reference database, irrespective of whether the individual has given its consent or not*" [21]. For the scope of this project it is important to state that our system does not permanently stores any data of our users without their consent or request.

1.3 Thesis Contributions

From the work reported in this thesis, the following contributions have made.
The developed source code is available at `https://github.com/sofiafernandescd/EmoReA`
The following paper was been published:
Sofia Condesso, Artur Ferreira, and Nuno Leite, “User Emotion Recognition from Speech and Text Messages with Machine Learning Techniques", Portuguese Conference on Pattern Recognition (RECPAD), Aveiro, Portugal, October 2025.

1.4 Organization of the Thesis

This document begins with Chapter 1, the current chapter, where we stated the motivation of this work and formalized the research question that we aim to respond and the proposed solution that we will be developing and testing for this problem.
Chapter 2 presents the literature review with an in-depth study of previous work on emotion recognition, starting with the theory of emotion, principal emotion recognition datasets and an overview of techniques used in all the phases of the machine learning cycle found present on the majority of the solutions presented in the literature.
Chapter 3 describes the methodology used, namely, the steps for data collection, data preprocessing, feature extraction, feature selection, modeling, training, and evaluating phases.
Chapter 4 presents the experimental results, analyzes and discusses the performance metrics. It also identifies limitations and elaborates on the discussion of the obtained results.
Chapter 5 draws conclusions, highlights key takeaways and findings, and suggests future work and potential points of improvement.
Finally, the Appendices provide complementary material that supports the main text. Appendix A details the dataset sources used throughout this research, including their characteristics, emotional taxonomies, and partitioning strategies. Appendix B describes the GeMAPS acoustic feature set, outlining its structure, feature categories, and relevance for emotion recognition tasks.

Chapter 2

Background and State of the Art

This chapter covers existing research on emotion recognition in multimedia content, including the problems of TER, SER, FER, and MER. Firstly, in Section 2.1, the theoretical foundations of emotion are explored, with discrete and dimensional models of emotion. Following this, unimodal and multimodal emotion recognition datasets are presented in Section 2.2, discussing data scarcity, cross-cultural variability, and ethical considerations. In Sections 2.3 and 2.4 the challenges and limitations of current research are discussed for unimodal and multimodal ER, respectively. The technical aspects of recognizing emotions in multimedia content are explored, including the types of preprocessing tools, feature extraction techniques, classification algorithms, and multimodal fusion approaches.

2.1 Emotion

The topic of emotions has been around since the 19th century, when Darwin (1859) theorized about the expression of emotion in animals, stating that emotions are biological traits that evolved through natural selection [22]. Some decades later, William James (1884) defined emotions as sensations of physical change, accompanied by physiological changes, such as facial expressions, muscle tension, and visceral activity [23].

Emotions are complex states that influence one's thought and behavior. At a physical level, physiological parameters such as heart rate, respiration cycles, and hormone levels can be measured to assess physiological arousal. In terms of the psychological aspect, cognition and behaviors can provide good information, through the analysis of voice, facial expressions, and body language.

Paul Ekman (1972) is a more recent pioneer in the field of emotions. He identified six basic emotions: anger, disgust, fear, happiness, sadness and surprise [24]. This categorization has been widely used in multidisciplinary research, such as psychology and affective computing.

2.1.1 Discrete Emotion Models

In discrete emotion theory, it is established that all humans have an innate set of basic emotions that are cross-culturally recognizable. Ekman's theory presents a discrete model of emotion that can be used with more or less granularity. Primary emotions, such as happiness, sadness, fear, anger, disgust, and surprise, were considered innate responses to stimuli, and believed to be universally experienced across cultures. Secondary emotions, such as guilt, shame, and envy, emerge from the interaction of primary emotions and social-cultural factors. These are often shaped by social norms, moral values, and interpersonal relationships.

Following Ekman's work, Plutchik's Wheel of Emotions (Figure 2.1) arranges emotions in concentric circles, where inner circles represent eight (instead of six) basic emotions, and outer circles are formed by blending the inner circle emotions. This is a hybrid model with both basic-complex categories and dimensional theories, which will be explored in the next section.

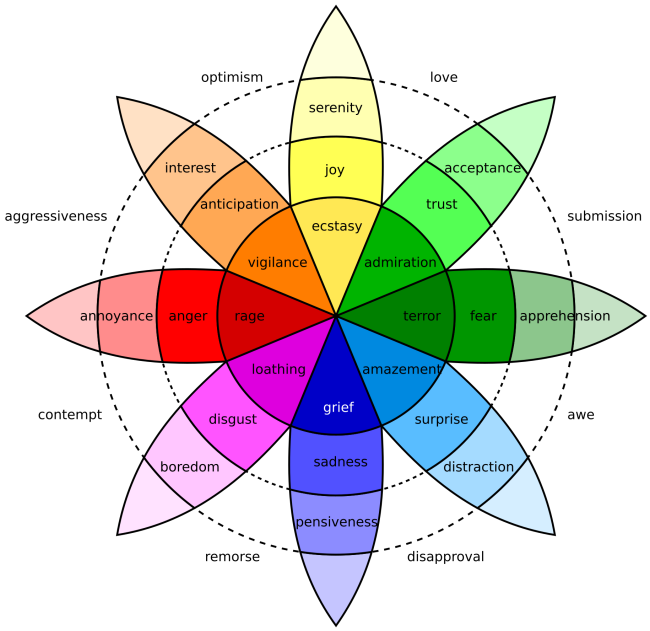


Figure 2.1 Plutchik's Wheel of Emotions. Image from [1].

In 1987, Shaver proposed a tree emotion model, which groups hierarchically emotions and allows for more or less granularity in the analysis, as can be observed in Figure 2.2.

In affective computing, namely, in multimodal emotion recognition datasets, emotion categories vary in number and designation, which can be a disadvantage in terms of uniformity of interpretation and practice. A strength of this paradigm is that we have a human recognizable concept that defines a specific emotion but that represents, simultaneously, a limit to the granularity of the analysis.

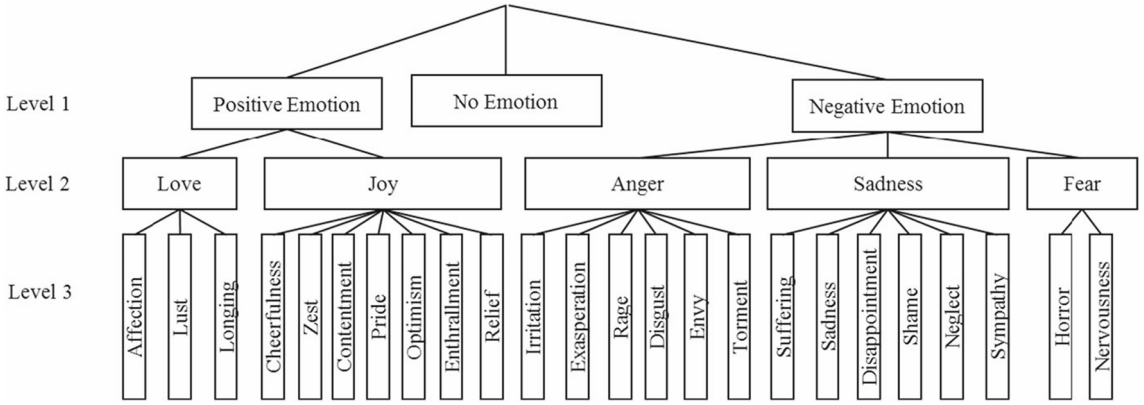


Figure 2.2 Shaver's emotion model. Image from [2].

2.1.2 Dimensional Emotion Models

Acknowledging the mixed and varying intensity of real-world emotions, dimensional models represent emotions along continuous dimensions. Wundt (1897) suggested dimensions such as "pleasurable versus unpleasurable" and "arousing or subduing". Schlosberg (1954) proposed "pleasantness-unpleasantness", "attention-rejection", and level of activation or "sleep-tension" [25]. Russell and Mehrabian (1974) further refined this, focusing on arousal, valence, and power.

Vector model - arousal and valence

Valence quantitatively represents the polarity of an emotion, positive or negative. Arousal quantifies the intensity of emotion, which can be higher, for example, excitement and rage, or lower on calmer emotions.

The vector model was first proposed by Bradley Greenwald, et al. in 1992. This model consists of an arousal dimension and a binary choice of valence that determines the direction of the vector and therefore defines different emotions using only arousal and valence [26]. One disadvantage is that the vector model holds that, at high arousal, the positive and negative valences are distinct from each other and hence neutral emotion cannot be well represented [26].

Positive emotions, such as joy, contentment, and gratitude, and negative emotions, such as fear, anger, and sadness, can be more or less intense. All these patterns can be computationally analyzed through a bi-dimensional representation of valence and arousal.

Circumplex model

Schlosberg's (1952) noted that emotions can be organized in a circle, represented by two bipolar dimensions. "*It is concluded that two widely used series of facial expressions can be described very well by locating them on a roughly oval surface whose longer axis is pleasantness-unpleasantness, the shorter axis being attention-rejection. Recognition of finer shades of emotion may depend on knowledge of the stimulus situation.*" [27].

James Russel (1980) proposed the circumplex model of emotion, that suggests that emotions are distributed in a two-dimensional circular space (Figure 2.3), containing arousal and valence dimensions, in the following order, with approximated angles: pleasure (0°), excitement (45°), arousal (90°), distress (135°), displeasure (180°), depression (225°), sleepiness (270°), and relaxation (315°) [3].

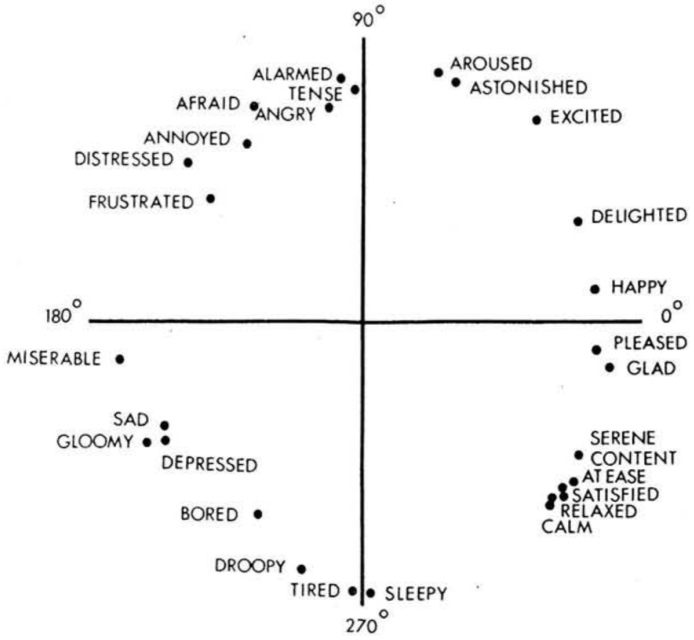


Figure 2.3 Direct circular scaling coordinates for 28 affect words. Image from [3].

The Positive Activation Negative Activation (PANA) model, created by Watson and Tellegen (1985) can be seen as a 45-degree rotation of the circumplex model, but some authors consider this model to be more similar to the vector model because the axes are unipolar [28].

2.2 Datasets

This Section delves into the characteristics of unimodal TER (Section 2.2.1), SER (Section 2.2.2) and FER datasets (Section 2.2.3), then it explores multimodal datasets (Section 2.2.4) and,

lastly, in the challenges in data collection and annotation (Section 2.2.5). The construction and characteristics of emotion recognition datasets significantly influence model performance. Speech and multimodal datasets, in particular, are categorized according to the manner in which emotional expressions are captured: acted, elicited, or natural.

Acted or simulated datasets are characterized by their professionally deliberate emotional expression [29, 30]. This method is considered the easiest approach to generate emotion-labeled speech datasets, accounting for approximately 60% of existing speech databases in a SER review [31].

Elicited or induced datasets are created by inducing emotional responses in participants through artificial situations, often without their knowledge [31]. This approach aims to capture more naturalistic emotional expressions compared to acted datasets, however, it raises ethical concerns regarding participant consent, as participants should be informed of their recording for research purposes [31]. Elicited speech represents an emotional middle ground, not neutral, and not artificially simulated [29].

Natural datasets convey spontaneous speech where the emotions are genuine [29]. Natural datasets consist of recordings of real-world scenarios, such as general conversations or call center interactions [31].

In the following sections, some of the most used datsets are presented for all covered tasks. The majority of the datasets were loaded directly to Python using *kagglehub* [32] library, therefore we also present dataset sources from Kaggle in Appendix A.

2.2.1 Textual Sentiment Analysis and Emotion Recognition Datasets

Textual datasets for sentiment and emotion analysis are crucial for developing models that understand emotional tone in written language. These datasets vary significantly in size, domain, and annotation granularity. Common types include social media data, product/movie reviews, news and dialogue data. Data can be annotated with sentiment polarity (negative, positive and neutral) or more granular emotion labels. A description of some datasets is provided in the following:

- **ISEAR** [33] dataset consists of 7666 sentences annotated with seven distinct emotion labels (joy, anger, sadness, shame, guilt, disgust, and fear). The dataset was constructed by Scherer et al. through cross-cultural questionnaire studies to 1096 people in 37 countries. This dataset comes from the answers of 1096 people from a wide range of cultural backgrounds. It has a balanced class distribution across its emotion labels, which can be used for generalized predictive inferences.

- **GoEmotions** [34] is a fine-grained TER corpus of 58009 curated comments extracted from Reddit. The dataset was created by Google Research Team, and it is richly annotated with 27 emotion categories, plus the neutral category. The authors provide the mapping of the 27 emotions to the six basic Ekman's emotions, which is very useful for our research.

  This corpus comprises 43410, 5427, and 5426 examples in the training, test and validation datasets, respectively. The original repository is hosted in GitHub [35], but the data can also be downloaded from Kaggle.

- **EmoryNLP** [36] is a corpus proposed by Sayyed M. Zahiri and Jinho D. Choi, in 2018. Each utterance is annotated with one of these seven emotions: sad, mad, scared, powerful, peaceful, joyful, and neutral. To measure the agreement between annotators, the authors computed Cohen's kappa and Fleiss' kappa scores. The corpus comprises 12606 utterances of "Friends" television show, taken from 897 scenes (4 seasons, 97 episodes), with 9934 utterances for training, 1344 for development and 1328 for test, all the sets exhibits class imbalance. The current state-of-the-art on EmoryNLP is CKERC [37], achieving 42.08% of weighted F1-score.

- **DailyDialog** [38] is a corpus built by Yanran Li et al. in 2017, comprising 13118 multi-turn dialogues. The utterances were labeled with one of the six basic emotions. The average number of tokens per utterance is 14.6.

- In 2018, Elvis Saravia et. al proposed an emotion dataset of **English Twitter messages** [39], for emotion recognition tasks, with six basic emotions: anger, fear, joy, love, sadness, and surprise.

- **Sentiment140** [40] is a well known sentiment analysis corpus, collected by Alec Go et al., in 2009. It comprises 1.6 million tweets extracted using the Twitter API, labeled with a polarity score from 0 to 4, where 0 is negative and 4 is positive.

2.2.2 Speech Emotion Recognition Datasets

SER datasets provide audio recordings annotated with emotional labels. Some multimodal datasets are vastly used to evaluate SER solutions. RAVDESS and IEMOCAP are examples of multimodal datasets that are commonly used to evaluate SER, and will be presented in Section 2.2.4.

- **EmoDB (Berlin Database of Emotional Speech)** [41] is a German language dataset

of acted emotional speech, recorded at the Technical University of Berlin with ten actors (5 female and 5 male). It comprises about 800 sentences, featuring seven emotions (anger, boredom, anxiety, happiness, sadness, disgust, and neutral). Reported performances vary depending on the experimental setup, with accuracies as high as 96% in some works. However, results are sensitive to factors such as feature choice, class balance, and evaluation protocol, given the relatively small size of the dataset.

- **TESS (Toronto Emotional Speech Set)** [42, 43] comprises 2800 recordings of two female actresses, aged 25 and 64 years, portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). This dataset has reached 99% accuracy.

Some datasets are designed for speaker-dependent SER. The main difference is that it contains speech from multiple speakers on the same input. Although in this thesis we will not develop a speaker-dependent solution, we can still use utterances from this kind of datasets.

2.2.3 Facial Emotion Recognition Datasets

FER datasets consist of images or videos of faces displaying emotional expressions. Common datasets include:

- **CK+ (Extended Cohn-Kanade)** [44] is a dataset of posed facial expressions, recorded with 123 participants, consisting of 593 video recording where peaked emotional expressions where taken as a dataset image. This resulted in 327 labeled examples, and it is considered a benchmark for FER research. State-of-the-art models achieve very high accuracies, often near 99% under identity-known splits.
- **FER2013 (Facial Expression Recognition 2013)** is a dataset of images with facial expressions organized by the seven emotions tackled in this work. This dataset was created by Kaggle for the FER Challenge, and consists of 28709 train images and 3589 test images with 48x48 of resolution in grayscale. Images are collected from the internet, often used for deep learning models. Novel works achieve 90-92%, but more modest (75%-80%) using seven emotions.
- **RAF-DB (Real-world Affective Faces Database)** [45] is a dataset of 29673 images of 100x100 pixels with grayscale and RGB colors, with skewed distributions in race, age, and expression. The dataset comprises the seven basic emotions, but also 12 compound emotions, and it is based on crowd-sourcing annotation. Accuracies range from 91% to 92% for seven basic emotions.

- **AffectNet** [46] dataset created in 2017, comprising the seven basic emotions plus the emotion contempt, which we filter when experimenting with it. The dataset is significantly more complex and diverse, has lower yet still meaningful top performance, generally in the 60%-65% range for 7 or 8-class settings under standard splits.
- **JAFFE (Japanese Female Facial Expressions)** [47, 48] comprises 210 images of ten Japanese women, in frontal position, presenting the following emotions: happiness, sadness, surprise, anger, disgust, fear, and neutral.
- **KDEF (Karolinska Directed Emotional Faces)** [49] is a dataset of posed facial expressions, recorded in 1998, with 70 amateur actors (35 female and 35 male), displaying the seven basic emotions.

The datasets CK+ and JAFFE have been the most commonly used face image datasets. In addition, FER2013, RAF-DB, and AffectNet have also been used in many studies.

2.2.4 Multimodal Emotion Recognition Datasets

MER datasets combine multiple modalities, such as audio, video, and text, to capture richer emotional information. Common datasets include:

- **IEMOCAP (Interactive Emotional Dyadic Motion Capture)** [50] is a multimodal dataset containing dyadic interactions with emotional labels, including scripted and improvised scenarios. Performances achieve 86.6%.
- **MELD (Multimodal Emotion Lines Dataset)** [51, 52] comprises dialogues from the TV series "Friends," annotated with emotion labels. Accuracies range from 60% to 65%, much lower than IEMOCAP due to conversational complexity and multi-speaker context.
- **SAVEE (Surrey Audio-Visual Expressed Emotion)** [53] consists of native English male speakers, aged from 27 to 31 years, that perform 480 utterances displaying the seven basic emotions. Multimodal approaches reach near 94%.
- **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)** [54] is a multimodal dataset containing audio and video recordings of emotional speech and songs, labeled with various emotions and intensities. Speech-only achives an accuracy around 80–85%, while video-only and multimodal often get much higher numbers (>90%).
- **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)** [55] is a set of 7442 original clips from 91 actors (48 male and 43 female), between the ages of 20

and 74, from a variety of ethnicities (african american, asian, caucasian, hispanic, and unspecified). The sentences were presented using one of six different emotions (anger, disgust, fear, happy, sad, and neutral) and four different emotion levels (low, medium, high, and unspecified). Research achieved 83–87% accuracy on multimodal approaches and around 75% using only speech.

- **CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity)** [56] is a large-scale multimodal dataset of video clips with sentiment and emotion annotations. The existing methods achieve a weighted F1 score around 80%.

According to the analysis of the datasets used in the past decade, EMO-DB, RAVDESS, and IEMOCAP have been the preferred choices for model testing [57, 58]. Table 2.1 summarizes the datasets considered in this study.

Table 2.1 Summary of emotion recognition datasets and best performance by modality. The modalities comprise *text* (T), *speech* (S) and *face* (F).

Dataset	Year	Samples	Modality	Performance	Notes / Domain
ISEAR	1994	7,666 sentences	T		Questionnaire-based, 7 emotions, cross-cultural
GoEmotions	2020	58,009 Reddit comments	T	0.51 Macro F1	27 emotions + neutral
EmoryNLP	2018	12,606 utterances	T		Friends TV dialogues, 7 emotions
DailyDialog	2017	13,118 dialogues	T		Multi-turn dialogues, 6 emotions
Twitter (Sravial)	2018	$\sim$20k tweets	T		English tweets, 6 emotions
Sentiment140	2009	1.6M tweets	T		Sentiment polarity (0–4)
EmoDB	2005	535 utterances	S	95.42%	German acted speech, 7 emotions
TESS	2009	2,800 recordings	S	99%	2 female actors, 7 emotions
CK+	2010	593 video sequences	F	99.68%	Posed expressions, benchmark dataset
FER2013	2013	28,709 train / 3,589 test	F	82.47%	Grayscale faces, 7 emotions
RAF-DB	2017	29,673 images	F	92%	7 basic + 12 compound emotions
AffectNet	2017	1M images	F	60%-65%	8 emotions (filtered to 7)
JAFFE	1998	210 images	F	97%-99%	Japanese females, 7 emotions
KDEF	1998	4900 images	F	92%-96%	70 actors, 7 emotions
IEMOCAP	2008	12h dyadic sessions	T,S,F	86.6%	Audio + video + transcripts, benchmark dataset
MELD	2019	13,000 utterances	T,S,F	60%-65%	Friends dialogues, multimodal
SAVEE	2007	480 utterances	T,S,F	93.22%	English male speakers, 7 emotions
RAVDESS	2018	7,356 clips	T,S,F	95.46%	Speech + song, acted emotions
CREMA-D	2014	7,442 clips	T,S,F	83%-87%	Crowd-acted speech + video
CMU-MOSEI	2018	23,500 videos	T,S,F	80% Weighted F1	Sentiment + emotion annotations

2.2.5 Challenges in Data Collection and Annotation

Data collection and annotation for emotion recognition pose several challenges, one of them being the scarcity of high-quality labeled emotional data, especially for certain modalities or emotional categories. The differences between acted and spontaneous emotion expression are significant, representing a challenge for the generalization of models. Natural datasets offer

the most realistic representation of emotional expression, but these datasets present significant challenges in emotion recognition due to their inherent variability [31].
Emotional expressions can be subjective and carry cross-cultural differences, leading to variations in annotations among different individuals from different cultures, therefore, potentially introducing bias into datasets [9]. Emotions are also highly context dependent, and without adequate contextual information, the data can be hard to use. Moreover, some emotional expressions are ambiguous or mixed, making accurate annotation difficult. When collecting and using emotional data some ethical concerns must be raised regarding privacy and consent, as discussed in the previous chapter.

2.3 Unimodal Emotion Recognition

This section reviews the state of the art in unimodal emotion recognition, considering each modality across four key dimensions: (*i*) preprocessing and tools for data handling and signal preparation; (*ii*) feature extraction and selection techniques that capture meaningful emotional information; (*iii*) classification methods, including both traditional machine learning and deep learning approaches; and (*iv*) related works with publicly available GitHub repositories supporting reproducibility.
In Section 2.3.1, we examine Textual Emotion Recognition (TER) and Sentiment Analysis (SA), highlighting linguistic preprocessing, embedding-based representations, and Transformer-based classifiers. Section 2.3.2 addresses Speech Emotion Recognition (SER), focusing on acoustic and prosodic features and classifiers ranging from traditional Machine Learning (ML) to neural networks. Finally, Section 2.3.3 considers Facial Emotion Recognition (FER), describing visual preprocessing pipelines, tools and frameworks for face detection and FER, and relevant computer vision implementations. These unimodal analyzes establish the methodological foundations and performance benchmarks that inform and motivate the multimodal fusion strategies discussed in Section 2.4.

2.3.1 Textual Emotion Recognition (TER) and Sentiment Analysis (SA)

While acoustic and visual modalities focus on how emotions are expressed, textual analysis deciphers what is communicated. Emotion recognition (ER) diverges from sentiment analysis (SA) by aiming to identify discrete emotions (for example, happiness, and fear) rather than polarity (positive/negative). In the textual modality in particular, SA is commonly performed instead of ER, due to the fact that the content from a text, can be spoken using different tones of voice and facial expressions, and therefore the exact expressed emotion can be harder

to recognize in comparison with the polarity of the content. ER is generally more expensive and complicated by linguistic nuances such as sarcasm and contextual dependencies, which demand sophisticated modeling approaches. There are essentially two main categories of approaches for this problem:

1. **Lexical** or **dictionary-based** approaches, which require a hand-crafted dictionary that associates words to sentiment polarity or emotions. By calculating statistics or using a rule-based approach, sentences or documents can then be rudimentarily classified, employing lexicons of emotional words [59]. One advantage of this method is that does not necessarily need much computation resources, but we have the model knowledge base limited to the dictionary.

2. **Model-based** approaches are more utilized in real world problems. Specifically neural networks have shown state-of-the-art results for the task of TER and Text Sentiment Analysis (TSA), with the advent of word embeddings.

When classifying emotions from text, one of the challenges that immediately arise is the cross-lingual aspect. A model capable to perform well in multiple languages is a model that most likely has a large number of parameters and was trained in huge amounts of public data, mainly from the Web which provide cross-lingual and cross-corpora information to those models. Due to this kind of models being more expensive to train and maintain, there are few emotion recognition studies in this field [9].

Preprocessing and Tools

Text preparation or text preprocessing aims to standardize input for downstream tasks. For traditional ML models it is common to perform the majority of the following steps, whilst with attention models and LLM ignoring some of these steps can output better results. The most common preprocessing steps are as follows:

- **Normalization** is the process of lowercasing, removing punctuation and expanding contractions (for example, “don’t” is normalized to “do not”). The lowercasing step and the removal of punctuation can harm the performance of larger models, by omitting capital letters and punctuation which can offer valuable information for TSA and TER [60].

- **Noise removal** filters non-linguistic elements such as URL and emojis, preserving sentiment-bearing symbols (for example, “!!!” for excitement) [60].

- **Stemming** and **lemmatization** are Natural Language Processing (NLP) foundational techniques that reduce words to their base or root form. Stemming refers to the heuristic

process of reducing words to their root form by removing inflectional affixes, often resulting in linguistically invalid stems (for example, converting “running” to “run” or “argued” to “argu”) [61]. In contrast, lemmatization employs vocabulary-based morphological analysis to derive the canonical dictionary form (lemma) of a word, preserving semantic validity (for example, mapping “better” to “good” or “is” to “be”) [60].

Tools like NLTK [62] and spaCy [63], using pre-trained language models, automate these processes and TER.

Tokenization is the process of segmenting raw text into interpretable units and it is indispensable in modern NLP. For transformer models, tokenization strategies must balance linguistic nuance, computational efficiency, and adaptability to diverse languages and domains.

The main word-level tokenization method is whitespace splitting, effective for English but inadequate for agglutinative languages. Traditional whitespace tokenization, though fast and simple, fails for languages lacking clear word boundaries, such as Turkish, where a single word like "çekoslovakyalılaştıramadıklarımızdanmışsınız" ("you are among those we couldn’t Czechoslovakianize") encodes a full clause [64]. Subword-level tokenization methods include:

- **Byte-Pair Encoding (BPE)** [65], popularized by models like GPT-3, iteratively merges frequent character pairs to form subword units. For example, the word “happiness” could decompose into ["happi", "ness"], allowing the model to recognize its root. While BPE effectively reduces Out-Of-Vocabulary (OOV) rates, it risks splitting meaningful parts of a word. For instance, "unhappily" could fragment into ["un", "happ", "ily"], obscuring the relationship between "un-" (negation prefixe) and "-ly" (adverbial suffix).

- **WordPiece** [66], used in BERT, refines BPE by prioritizing token pairs that maximize language model likelihood. This method preserves common prefixes and suffixes, such as retaining "##ing" for verb forms (for example, "running" is decomposed in ["run", "##ning"]). However, WordPiece requires domain-specific retraining to handle specialized vocabulary, such as medical terms or slang in social media texts, a limitation for emotion recognition in niche contexts. Subword methods like BPE and WordPiece handle rare/misspelled words by breaking text into statistically frequent units, but at the cost of fragmenting semantic coherence. For example, the word "embeddings" might split into ["em", "##bed", "##ding", "##s"], diluting its connection to the root "embed".

- **SentencePiece** [67], employed in T5 and ALBERT, operates independently of whitespace, making it ideal for agglutinative languages like Finnish or scripts like Chinese. While versatile, SentencePiece may over-segment short words, inflating token counts unneces-

sarily.

- The **Unigram Language Model** [68], used in XLM-R [69], probabilistically prunes subwords to optimize a vocabulary. This method retains morphological structure—decomposing "running" into ["run", "##ning"], but demands careful tuning to avoid under-segmenting rare words. SentencePiece and Unigram excel in multilingual settings but struggle with domain adaptation. A model trained on formal news corpora might missegment informal phrases like "OMG!!! Sooooo cool!!!" into ["_OM", "G", "!!!", "_So", "oooo", "_cool", "!!!"], losing the emotional emphasis conveyed by elongated vowels and repeated punctuation.

Frameworks like Hugging Face Tokenizers unify BPE, WordPiece, and SentencePiece under a single API, enabling seamless integration with Transformer architectures [70]. For low-resource languages, tools such as spaCy's rule-based tokenizers extend transformer pipelines to prioritize morphological richness. In modern NLP, tokenization can be seen as the starting point of feature extraction, because tokens are mapped to numerical representations.

Feature Extraction and Selection

When working with text in machine learning, feature representation is composed mainly by two categories: Bag-of-Words (BoW) and embedding methods.
BoW methods represent texts as vectors in which each position corresponds to a word in the vocabulary. The entries may be binary (indicating presence or absence) or frequency-based (indicating the number of occurrences of each word in a document). A widely used refinement of this approach is Term Frequency–Inverse Document Frequency (TF-IDF), which assigns a weight to each word based on its importance in the document relative to its prevalence across the corpus.
Beyond BoW, topic modeling methods such as Latent Dirichlet Allocation (LDA) [71] and Non-Negative Matrix Factorization (NMF) [72] can uncover latent semantic structures and are commonly applied to tasks such as document clustering and recommendation.
Embedding methods provide an alternative to these sparse representations. An embedding is essentially a dense, machine-understandable vector of numbers that captures the essence of the source information. Static embeddings such as Word2Vec [73] and GloVe [74] map words to dense vectors, where semantic similarity is captured by geometric proximity (for example, the vectors for "joy" and "delight" should be close in space). However, such models do not disambiguate polysemous words. Contextual embeddings, introduced with models such as BERT [75] and RoBERTa [76], produce dynamic representations that account for context, allowing

different meanings of a word such as “crash” to be correctly distinguished in financial versus automotive domains.

Classification Methods

In early studies, TSA was more frequent and researchers would classify if a text showed positive, negative, or neutral sentiment. With the advent of LLM, the ability of these systems to understand emotions in texts improved, and nowadays we can achieve state-of-the-art performance by fine-tuning some of these models. Figure 2.4 shows a timeline overview of SER approaches.

Linguistic inquiry and word count (LIWC) was first developed by James W. Pennebaker et al., and the first major version, LIWC2001, was published [77] in a manual detailing its development and psychometric properties. It originated from the need for a reliable way to analyze the psychological content of essays. A simple computer program to count a small number of emotion-related words, which eventually evolved into the LIWC software. The National Research Council of Canada Emotion Lexicon (also known as EmoLex) [78] is a freely available list of English words that are associated with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (positive and negative).

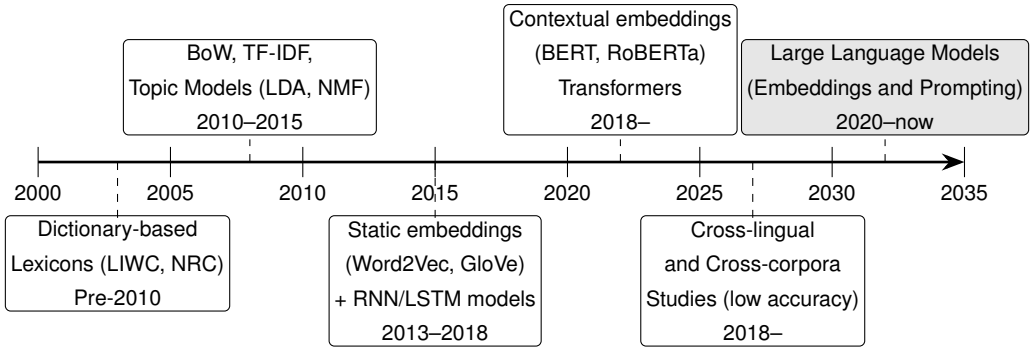


Figure 2.4 Timeline of Text-Based Emotion Recognition (TER) and Sentiment Analysis (SA), from dictionary-based approaches to modern Transformer-based models.

TSA is an easier task, when compared to TER, and basic ML models like Logistic Regression, Naive Bayes and Support Vector Machines (SVM) can achieve good performance on classifying the sentiment as positive, negative, or neutral. Decision Trees and Random Forests are used for both TSA and TER, normally using BoW methods or static word embeddings.

Specifically for the problem of TER, the study [59] compared traditional ML and deep learning for ISEAR dataset. They obtained 64% of accuracy, when using BERT embeddings, classified by an SVM. Furthermore, they tested neural networks with different methods for feature extraction, where TF-IDF got the higher accuracy of 58%. We have the following approaches:

- **Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection** [79], proposed by Zhu et al., consists of a topic-augmented language model with an additional layer specialized for topic detection. The topic-augmented language model is then combined with commonsense statements derived from a knowledge base based on the dialogue contextual information.

  `https://github.com/somethingx678/TodKat`

- **EmoBERTa** [80] models, proposed by Kim and Vossen in 2021, are based on RoBERTa-base and RoBERTa-large, trained on both MELD and IEMOCAP datasets. This approach enriches Transformer models by including the speaker identity in the sequence information, and achieved 65.61% accuracy on MELD dataset.

  `https://github.com/tae898/erc`

- **Twitter Emotion Recognition using RNN** is implemented, an embedding layer, followed by two bidirectional LSTM layers with 20 units each.

  `https://github.com/katoch99/Twitter-Emotion-Recognition`

TER and TSA Using AI Conversational Agents Powered by LLM

Recently, LLM have been reshaping the field of ML, due to their natural language processing and understanding, triggering advances in text classification, question answering, and lately text generation, where we can use some of them as conversational agents. In this kind of approach, the preprocessing step is not considered because LLM are pre-trained on extensive datasets and can understand nuanced meanings within natural text. Figure 2.5 names some of the most important models over time.

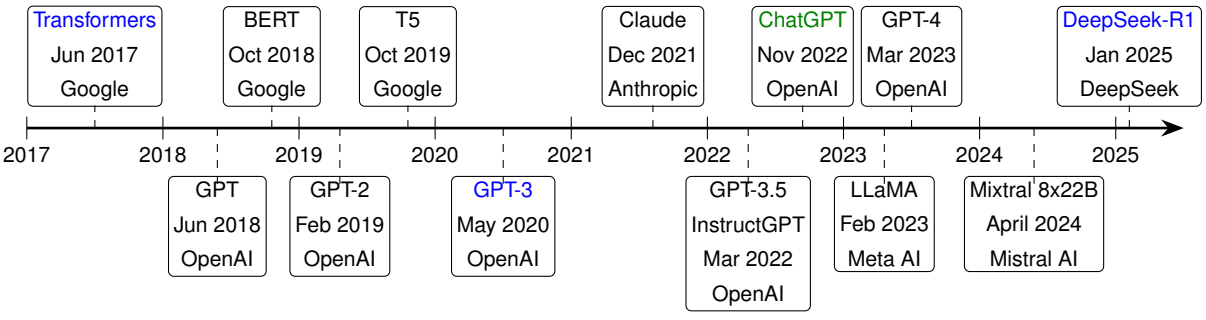


Figure 2.5 A brief history of LLM .

Retrieve Augmented Generation (RAG) is a technique that has been increasingly used on real-world applications, because it has the ability to deal with dynamic data, improving the capability of a model through retrieving up-to-date external corpus of information (in various formats). It

consists of augmenting the prompt that was given to the model and generating an answer that uses that context and information.

Model specialization can be done through fine tuning and prompt engineering. Fine-tuning is to adjust the weights of a pre-trained language model based on a specific domain, teaching the model to perform better at a specific task. To avoid retraining the models for each dataset, it is common to leverage the general language understanding of LLM and make use of them without any changes on their architecture, usually trough zero-shot and few-shot prompting. An example of zero-shot prompting is as follows:

Given the following dialogue identify the underlying emotion: [text]

Answer with only one of: [happy, disgust, sad, angry, fear, surprise, neutral]

In few-shot prompting, we include labeled examples of the task in the prompt, to help the LLM to better understand:

Classify the emotion in the following dialogues:

Example 1: I can't believe he forgot my birthday. Emotion: sad

Example 2: You never listen to me! Emotion: angry

Now classify: I didn't expect that at all! Emotion: _____

Another way querying the LLM about the emotions present on a text (based on [81]):

"Can you assign a score between 0 and 1 to each of these emotions based on what is expressed in the text: happy, disgust, sad, angry, fear, surprise and neutral?"

Running large language models locally is particularly important in the context of TER. First, it reduces latency, since the requests do not need to travel to external servers, which is crucial when working with real-time or interactive applications. Second, it ensures data privacy, as sensitive text data never leaves the user's machine. Frameworks such as Ollama allow deployment and execution of LLM locally, making them practical for research and privacy-preserving applications in emotion recognition.

This section is dedicated to highlighting some experiments with LLM, particularly in the task of emotion recognition. Most of the time, researchers have tackled the problem of Emotion Recognition in Conversations (ERC), but solutions are evaluated in popular ER datasets, such as IEMOCAP. The following works with github repositories are relevant:

- **ACL 2024 WASSA Task 2** implementation achieved 1st place out of 72 teams in the shared task on cross-lingual emotion detection. The proposed ensemble approach strategically combines the strengths of multiple models across languages, reaching 63% F1-score, 65% precision, and 61% recall, demonstrating robust performance in multilingual scenarios.

https://github.com/1024-m/ACL-2024-WASSA-TASK-2/

- **InstructERC** [82] introduces an instruction-tuned approach for ERC, where large language models are guided by task-specific prompts. This method aligns LLM outputs with emotion classification objectives, allowing better adaptation to conversation datasets and improved interpretability.
  https://github.com/LIN-SHANG/InstructERC

- **LaERC-S: Improving LLM-based Emotion Recognition in Conversation with Speaker Characteristics** [83], proposes incorporating speaker-specific information into the prompt design of LLM for conversation-based emotion recognition. By modeling speaker traits and conversational dynamics, the approach enhances the ability of LLM to disambiguate emotions across dialogue turns.
  https://github.com/lin-shang/LaERC-S

2.3.2 Speech Emotion Recognition (SER)

The recognition of emotions from speech has been a subject of research since the mid-1980s, when statistical properties of acoustic features were first analyzed to infer the emotional state of a speaker. Early works, such as those by Van Bezooijen (1984) and Ververidis and Kotropolos, primarily focused on extracting statistical properties from prosodic and spectral patterns, to infer the speaker's emotional state [29].
Speech conveys affective information through two complementary channels: explicitly, via the linguistic content of the spoken message, and implicitly, through paralinguistic cues such as prosody, intensity, and spectral characteristics [84]. While the former was addressed in Section 2.3.1, the latter constitutes the focus of Speech Emotion Recognition (SER). Despite decades of study, decoding these paralinguistic signals remains challenging. Humans can generally recognize the six basic emotions with reasonable accuracy, and machine learning models have increasingly approached similar performance levels in benchmark datasets. Nevertheless, issues such as inter-speaker variability, noise robustness, and contextual dependencies (for example, correlations between sequential utterances in a dialogue) continue to pose open research challenges [85].

Preprocessing and Tools

Signal processing is the first step of SER pipeline. In digital audio processing, an audio signal is a continuous waveform, that must be discretized into samples for computational analysis.

These samples represent the amplitude of the audio signal at specific points in time. Speech signals are often subject to channel distortion, background noise, and inter-speaker variability, making it necessary to normalize and enhance the signal before feature extraction [86]. The goal of preprocessing is to preserve discriminative emotional cues while mitigating irrelevant variability introduced by recording conditions or individual speaker traits [87].

The raw waveform is typically segmented into short overlapping frames (example, 20–40 ms) to capture quasi-stationary properties of speech. Frame overlap ensures temporal continuity, reduces artifacts at frame boundaries, and preserves information that would otherwise be lost in non-overlapping segmentation.

Windowing functions, like Hamming and Von Hann windows, are applied to minimize spectral leakage. Additional preprocessing steps often include denoising, normalization, and Voice Activity Detection (VAD), which helps to discard silent or irrelevant segments [87]. The duration of the segment is another critical design choice. Most SER systems analyze speech in segments of 3–5 seconds [88], and Interspeech 2020 recommended 4.2 seconds as an optimal input length for dimensional emotion recognition tasks.

Several open-source toolkits facilitate the extraction of acoustic features. *openSMILE* [89] remains the most widely adopted in SER research due to its standardized feature sets, such as Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). Other commonly used libraries include *librosa* [90], popular for spectral and prosodic feature extraction in Python; *Praat* [91], widely used for phonetic and prosodic analyzes; and *openEAR* [92], which extends *openSMILE* with additional emotion recognition utilities. These tools aim to extract as much affect-relevant information as possible while minimizing loss due to compression or abstraction [9].

Transcript extraction is usually performed in multimodal pipelines, where the audio signal and the transcript are analyzed to recognize emotions. Speech can be transcribed into text to enable emotion recognition from linguistic perspectives. Automatic Speech Recognition (ASR) models such as *Whisper*, *Vosk*, and the *SpeechRecognition* are examples of libraries to extract speech transcripts. Whisper, in particular, demonstrated high robustness to noise and multilingual adaptability.

Feature Extraction and Selection

Feature engineering in SER encompasses both the extraction of salient acoustic characteristics from the signal and the subsequent selection of the most informative features for classification. The input is a digital representation of the audio waveform, characterized by three fundamen-

tal aspects: amplitude (signal intensity), frequency (rate of oscillation), and waveform shape (signal pattern). From these basic properties, more complex descriptors can be derived.
Feature representation in SER is commonly organized across multiple levels of speech: *frame*, *segment*, and *utterance* levels [9]. This hierarchical approach captures both fine-grained temporal variations and longer-term contextual patterns:

- **Frame-level features** (short-term) describe instantaneous acoustic properties, typically computed over windows of tens of milliseconds. They are essential for capturing dynamic vocal fluctuations.
- **Segment and utterance-level features** (long-term) are statistical aggregations (for example, mean, standard deviation, and variance) of frame-level descriptors. These provide a higher-level view of prosodic and spectral trends, which improves robustness against local noise and variability [93].

To reduce redundancy and standardize feature representation, several curated sets have been proposed. The most prominent is the **GeMAPS feature set**, consisting of 18 Low-Level Descriptor (LLD) and their functionals (62 features in total). The extended version, **eGeMAPS**, expands this to 25 LLD and 88 features, offering a balance between compactness and discriminative power, and has become a *de facto* standard in affective computing [94].
The extracted features can be grouped into several categories: prosodic, spectral, and voice quality features.
Prosodic features are related to the rhythm, stress, and intonation of speech, which are crucial for emotion identification. Prosodic features include intensity, pitch (fundamental frequency), energy, duration, and speaking rate [9, 31]. **Intensity** refers to the loudness or amplitude of the speech signal. Variations in intensity can reflect changes in vocal effort and emotional arousal. **Pitch**, or fundamental frequency (F0), represents the perceived highness or lowness of the voice and intonation. Variations in pitch, such as rapid changes or sustained high or low pitch, can signal different emotional states. **Speaking rate** refers to the speed at which speech is produced. **Energy** features are related to the amplitude of the speech signal within a given time frame. Variations in energy can reflect changes in vocal effort and stress. These are fundamental features that often correlate with emotional arousal, as higher energy levels can indicate heightened emotional states like anger or excitement. **Teager energy operator (TEO)** features are more advanced descriptors that capture the nonlinear energy production in the vocal tract [87]. TEO is calculated in the time domain, and it is often used to track instantaneous energy changes. These features provide insights into vocal tract articulation,

capturing subtle changes like vocal effort and tension that may be associated with emotional states.
Emotional states often modulate these properties: for instance, high intensity and pitch variability may signal anger or excitement, whereas reduced energy and slower speaking rates are associated with sadness.

Spectral features represent the frequency distribution of the speech signal, providing insights into the spectral envelope and vocal tract characteristics. These features, including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), and formants, represent the frequency distribution of the speech signal and help in understanding the resonance and timbre of the sound [9, 31].

- **MFCC** are a set of features that approximate human auditory perception. They are derived from the power spectrum of the speech signal, transformed using the Mel scale, which is non-linear, unlike linear Hertz scale. The human ear is more sensitive to subtle differences in lower frequencies than in higher frequencies and Mel scale mirrors this, with a more detailed, linear-like representation at lower frequencies and a more compressed, logarithmic-like representation at higher frequencies, emphasizing frequencies relevant to human hearing. A great number of works use MFCCs as features for the classification of emotions, which shows they are an adequate way of analyzing emotions compared to other commonly used speech features (for example, loudness, formants, linear predictive coefficients etc.) [95].

- **LPCC** are obtained by a technique that models the vocal tract and provides information about the spectral envelope. It represents the speech signal as a linear combination of past samples, capturing the resonant frequencies of the vocal tract.

- **Formants** are also considered spectral features, as they are extracted from the spectral representation of the speech signal. Formants are resonant frequencies of the vocal tract, determined by its shape and size. They provide information about vowel articulation, as different vowels are characterized by distinct formant patterns. Emotional states can influence vocal tract configuration, leading to subtle changes in formant frequencies and bandwidths. The first studies on emotional speech consisted of analyzing formants and durational characteristics [9].

Voice quality features describe the characteristics of the vocal source, providing insights into vocal fold vibration and articulation. These features, such as jitter, shimmer, and Harmonics-to-

Noise Ratio (HNR), describe the characteristics of the vocal source, as follows:

- **Jitter** refers to the variation in fundamental frequency from one vocal fold vibration cycle to the next. Increased jitter can indicate vocal fold instability, which may be associated with certain emotional states.
- **Shimmer** refers to the variation in amplitude from one vocal fold vibration cycle to the next. Increased shimmer can also indicate vocal fold instability.
- **HNR** measures the proportion of harmonic components to noise in the speech signal, reflecting vocal fold regularity. Higher HNR indicates a clearer, less noisy voice.

Voice quality can be described by laryngeal and supralaryngeal settings, as per Laver's approach [9].

The selection of acoustic feature has also been studied and a curated sets of features were proposed, like the GeMAPS features [94]. GeMAPS is a curated feature set that consists of 18 LLD and their functionals, that standardizes a powerful set of acoustic parameters for voice analysis, particularly in affective computing and paralinguistics, resulting in a total of 62 features. eGeMAPS is the extended GeMAPS set, which includes an additional 7 LLD, bringing the total number of LLD to 25. With functionals, the eGeMAPS set has 88 features.

Classification Methods

The classification task in SER involves mapping extracted acoustic features to emotional labels. Although several machine learning and deep learning techniques have been used, there is no universally accepted classifier, and aggregating classifiers is also a common strategy [96]. Figure 2.6 overviews the evolution of SER approaches regarding feaures and classifiers.

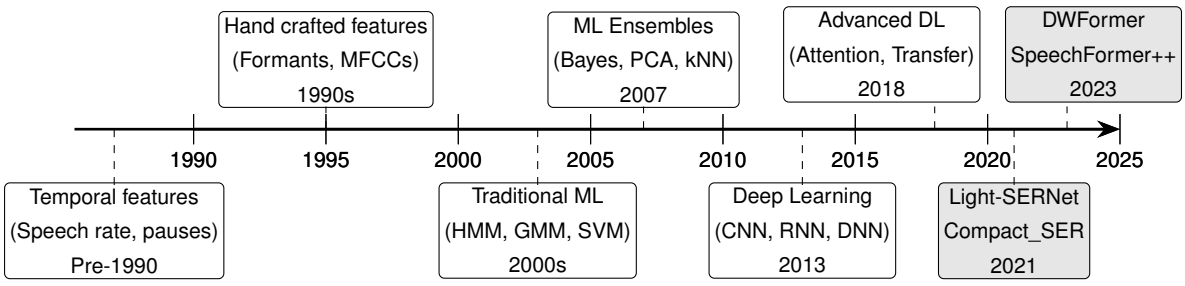


Figure 2.6 Timeline of Speech Emotion Recognition development from hand-crafted features in the 1990s to modern deep learning-based approaches.

Early methods using traditional machine learning approaches, such as Hidden Markov Models (HMM)[97, 98], Gaussian Mixture Models (GMM)[99], SVMs [100, 101], Decision Trees[102], K-Nearest Neighbours (KNN), and Naive Bayes [31, 93, 103] have been successfully employed

to classify from acoustic features, with SVM, HMM and GMM being particularly prevalent in literature [96]. Linear classifiers like Bayesian Networks and Maximum Likelihood Principle have been also employed [31] as well as k-means, Principal Component Analysis (PCA), Logistic Regression, Matrix Completion, Sparse Learning, and Multi-Graph Learning [103]. As noted, a variety of models can be applied to these problems, and other researchers tried to enhance performance by ensembling two or more methods referred to above.

Deep learning models require less preprocessing of the input and are able to perform feature learning and extraction through the train of the entire neural network architecture. With respect to SER, Deep Neural Networks (DNNs) such as Convolutional Neural Networks (CNN) [14, 104, 105], and Recurrent Neural Networks (RNN) [9, 87, 93], using Long Short-Term Memory (LSTM) [106] and Bidirectional Long Short-Term Memory (BiLSTM) [107], are increasingly used due to their capacity to learn complex patterns through time. With the ascension of Transformer architecture, more deep learning based enhancement techniques such as attention mechanism [108, 109], ensemble modeling [110, 111, 112], transfer learning [113, 114, 115], and adversarial learning [116, 117] are more complex techniques that have been employed.

An analysis of the datasets used in this past decade indicates that EMO-DB, RAVDESS, CASIA, and IEMOCAP datasets have been the most preferred choices for model testing [58].

The following works, related to this topic, are available:

- **Light-SERNet** [104] (2021) is an efficient and lightweight fully convolutional neural network designed for speech emotion recognition on systems with limited hardware resources. The model achieved 85.71% accuracy on EmoDB and between 65%–70% on IEMOCAP.

  `https://github.com/AryaAftab/LIGHT-SERNET`

- **Compact-SER** [118] (2021), proposed by Shirian and Guha, introduces a compact and scalable graph-based representation of speech data. Based on graph signal processing, where speech is modeled as a cycle or line graph, the method employs a GCN architecture that performs exact graph convolution, unlike the approximations in standard GCN. Reported accuracy ranges from 60% to 80% on EmoDB.

  `https://github.com/AmirSh15/Compact_SER`

- **Emotion Neural Transducer (ENT)** [119] (2024), proposed by Shen et al., addresses the challenge of modeling fine-grained emotion dynamics at the temporal level, a less explored area in SER research. ENT achieves 73.88% accuracy on IEMOCAP, comparable

Table 2.2 Summary of related works in speech emotion recognition classifying seven emotions.

Ref.	Year	Dataset	Feature Extraction	Model Architecture	Notes
Pan et al. [100]	2019	EmoDB and Chinese corpus	energy, pitch, MFCC, and MEDCs	SVM	91.3% on Chinese corpus and 95.1% on EmoDB.
Aftab et al. [104]	2021	IEMOCAP and EmoDB	three parallel CNN are applied to the MFCC to extract time and frequency features	Local feature learning blocks (LFLB)	While the model has a smaller size than that of the state-of-the-art models, it achieves 70.23% on IEMOCAP and 94.21% on EMODB.
Shen et al. [118]	2021	IEMOCAP and MSP-IMPROV	model speech signal as a cycle graph or a line graph	Graph Convolution Networks (GCN)-based architecture that can perform an accurate graph convolution	Achieve 65.29% on IEMOCAP with significantly fewer learnable parameters, indicating its applicability in resource-constrained devices.
Shen et al. [119]	2024	IEMOCAP and ZED	Extend typical neural transducer with emotion joint network to construct emotion lattice for fine-grained SER	Emotion Neural Transducer	ENT architecture achieved 72.43% WA on IEMOCAP, and FENT architecture excels at fine-grained SER on ZED regardless of lattice while ENT obtains better UA in typical utterance-level SER.

to state-of-the-art methods while requiring fewer parameters.

`https://github.com/ECNU-Cross-Innovation-Lab/ENT/`

- **DWFormer (Dynamic Window Transformer)** [120] is a Transformer-based model that dynamically adjusts the window size to better capture emotional cues in speech. It reports 72–73.9% accuracy on IEMOCAP and 48.5% on MELD.

  `https://github.com/scutcsq/DWFormer`

- **SpeechFormer++** [109] (2023) extends Transformer-based architectures for SER and

other paralinguistic tasks, focusing on efficiency and hierarchical feature modeling.
https://github.com/HappyColor/SpeechFormer2

- **TrustSER** [121] (2024) proposes a framework to evaluate SER systems beyond accuracy, focusing on privacy, safety, fairness, and sustainability. Built on pre-trained models (TERA, Wav2vec 2.0, WavLM, and Whisper), using raw wave or mel-spectrogram as features, it fine-tunes frozen encoders with weighted hidden-layer aggregation and CNN + dense layers. Reported Unweighted Average Recall (UAR) on IEMOCAP reaches 65–70%, while also benchmarking robustness against gender inference attacks and adversarial perturbations.
https://github.com/usc-sail/trust-ser

2.3.3 Facial Emotion Recognition (FER)

FER systems typically operate in three main stages:

1. Face detection - The system locates and isolates faces within an image or video frame. Preprocessing steps can be done to improve face detection.
2. Feature extraction - Key facial features such as eyebrows, eyes, mouth, and their relative positions are identified and extracted.
3. Emotion classification - After training an AI model under the same conditions, for predicting emotions, the extracted features are analyzed to predict the most likely emotion that is being expressed.

To classify facial expressions we have two main categories of modeling approaches. Static modeling is when the classifier looks at a single image or feature vector at a time, with no concept of temporal evolution, whereas dynamic modeling encompasses the representation of a sequence of frames and their transitions, over time [122].

Preprocessing and Tools

In FER, while in curated datasets preprocessing can be ignored, in real world applications it is a critical phase of FER, indispensable in removing noise, correcting distortions and enhancing salient features for downstream tasks [123]. Image processing steps include grayscale conversion, face detection, and alignment, which collectively standardize the input data for robust feature extraction [124].

Grayscale conversion is the process of reducing color images to a single luminance channel. Empirical studies suggest that color information often contributes minimally to FER accuracy while tripling processing demands [125]. The normalized grayscale formula is widely adopted and is derived from the CIE 1931 luminance weights, aligning with human photoreceptor sensitivity [124], computed by

$$Y = 0.2989\,R + 0.5870\,G + 0.1140\,B \tag{2.1}$$

Tools like OpenCV (via cv2.cvtColor) and Python Imaging Library (PIL) implement this conversion efficiently, often leveraging hardware acceleration for real-time applications.

Face and Landmarks Detection

Face detection localizes facial regions of interest (ROI), a prerequisite for subsequent analysis. Early approaches relied on geometric features like fiducial points and angles [126], while modern systems leverage appearance-based methods. Haar cascades [127] and histograms of oriented gradients (HOG) [128] dominated pre-deep-learning eras.

The Viola-Jones framework [127] revolutionized face detection by enabling real-time performance. It relies on Haar-like features, which use rectangular filters to encode edge and texture patterns, and the Integral Image, which allows rapid feature computation through precomputed pixel summations. AdaBoost is then applied to select the most discriminative features and train a cascade classifier that quickly rejects non-face regions. This method, implemented in OpenCV via `cv2.CascadeClassifier`, remains popular for embedded systems, although it can struggle with occlusions and extreme poses.The Histogram of Oriented Gradients (HOG) [128] extracts edge orientation histograms, combined with linear SVMs for classification. Dlib's HOG detector achieves high precision on frontal faces but lags in speed compared to deep learning methods.

Deep learning detectors like MTCNN (Multi-Task Cascaded CNN) [129] and RetinaFace [130] use CNN to jointly detect faces and landmarks. These achieve state-of-the-art accuracy in unconstrained environments (for example, occlusions, and profile views) at the cost of higher computational load. MTCNN is a pipeline of three CNN applied in stages:

1. P-Net (Proposal Network) that quickly scans the image at multiple scales, and proposes candidate face regions with bounding boxes and landmark offsets.

2. R-Net (Refine Network) which takes candidate boxes from P-Net and rejects false positives, refining bounding box locations.

3. O-Net (Output Network) which further refines detections. Outputs the final bounding box plus 5 facial landmarks (eyes, nose, and mouth corners).

Face Alignment

There are some preprocessing steps that should be used to improve the algorithm performance. Rotation correction is the first step, which consists of taking the center point between the eyes, and rotate the image to get the eyes horizontally aligned. Images frequently have information from the environment that is not related to the facial expression, so cropping the image to get only the face pixels is essential, since that information can decrease the accuracy. Downsampling aims to reduce image size and resolution while preserving essential visual characteristics. Lastly, intensity normalization of the image is also commonly perform when using FER algorithms. Face alignment normalizes facial geometry by mapping detected regions of interest to canonical coordinates. Faces in images can be alignable or non-alignable. Techniques include:

- Active Shape Models (ASM) [131] iteratively deform a statistical shape model to match fiducial landmarks (for example, eyes, and mouth). These methods require manual initialization and struggle with extreme expressions.
- Deep Alignment Network (DAN) [132] and HRNet [133] use cascaded CNN to predict landmarks robustly. Tools like Dlib (dlib.shape_predictor) and OpenFace provide pretrained models for 68-point landmark detection.
- 3D Morphable Models (3DMM) [134] is a frame that fits 3D face models to 2D images, resolving pose variations by estimating head rotation and translation.

Feature Extraction and Selection

Facial feature extraction reduces data dimensionality while preserving discriminative information. Traditional methods include:

- **Local Binary Pattern (LBP)** encode micro-texture patterns by comparing the center pixel value with its neighborhood intensities [135, 136]. OpenCV `cv2.LBPHFaceRecognizer` is a classical face recognizer that applies LBP histograms for real-time texture analysis, though sensitivity to lighting limits robustness. It can be adapted to FER although it does not scale well.
- **Local Directional Pattern (LDP)** [137, 138] feature is obtained by computing edge response values across eight distinct directions at each pixel location. A code is then generated based on the relative magnitudes of these responses. Consequently, each facial

image is represented as a collection of LDP codes, which serve as the basis for the recognition process.

- **Histogram of Oriented Gradients (HOG)** can also be used for further feature extraction, capturing micro-expression in edge orientation histograms from localized cells, in a way that presents robustness to illumination changes [128]. Normalized block descriptors enhance illumination invariance. Scikit-image `skimage.feature.hog` provides optimized implementations.

- **Gabor wavelets** are known for their ability to represent visual information, and have also been used as a feature extraction technique [139].

- **Scale-Invariant Feature Transform (SIFT)** detects keypoints invariant to scale and rotation using Difference-of-Gaussians pyramids [140]. VLFeat and OpenCV `cv2.SIFT_create` enable descriptor extraction, but computational cost limits real-time use.

- **Facial Action Coding System (FACS)** was originally created by Carl-Herman Hjortsjö with 23 facial motion units in 1970, it was subsequently developed further by Paul Ekman, and Wallace Friesen. It encodes facial muscle movements into Action Unit (AU). Modern tools like OpenFace [141, 142] automate AU detection via CNN, allowing further emotion classification based on rules for mapping AU to emotions. Ekman's description of the six emotions is linguistic and find universal associations between emotion descriptions and facial expressions in terms of, for example facial actions [126].

- **Feature Point (FP)** were standardized by datasets like Multi-PIE (2005)[143] FACS. Common markup schemes consist of 68-point markup (eyebrows, eyes, nose, mouth, and jawline) and dense landmarks (see Figure 2.7). Anchor points on a neutral face enable consistent alignment across poses/expressions, but occlusions or extreme poses can obscure FP, requiring robust detection (for example, using 3DMM or GAN)[4].

Paradigms shifted with the advent of deep learning, with CNN automating feature learning. Pretrained models like AlexNet (2012)[144], VGGNet (2014)[145], and ResNet (2015)[146] extract hierarchical features, outperforming handcrafted methods [147]. Attention mechanisms like Vision Transformers (ViT) [148] use self-attention to focus on salient regions (for example, eyebrows for surprise). Some hybrid approaches, like DeepEmotion, concatenate CNN features with FACS-coded AU, improving interpretability and accuracy on compound emotions [149].

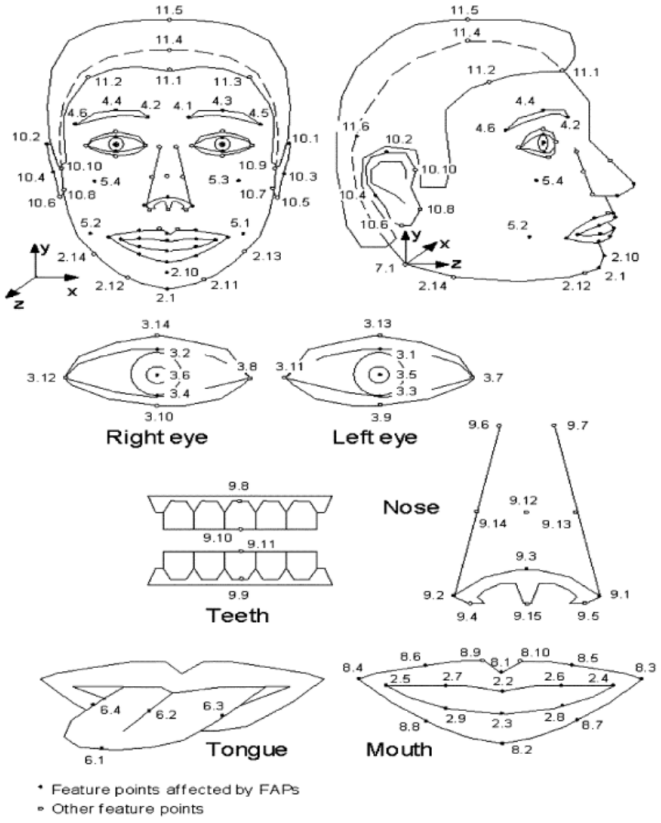


Fig. 3: The 84 Feature Points (FPs) defined on a neutral face. Figure reprinted from [28] [C].

Figure 2.7 Facial Feature Points. Image from [4]

Classification Methods

Similar to previous modalities, there is no one-fits-all model for FER. Traditional ML solutions usually extract features from face and landmarks, and classify static feature vectors without considering temporal information. HMM are a powerful traditional ML technique specifically designed for sequential data. They model a system with a set of hidden states and the probabilities of transitioning between them. Figure 2.8 shows a basic workflow using traditional ML.

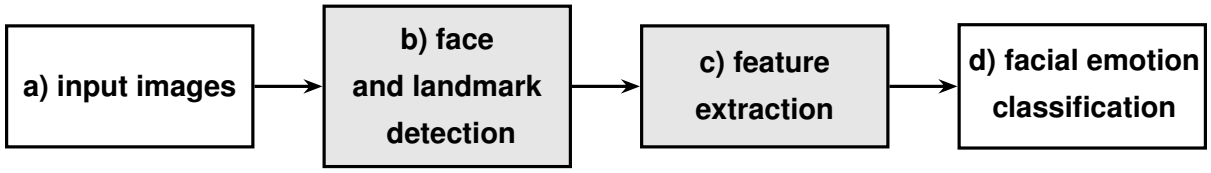


Figure 2.8 Typical FER workflow with traditional ML

On the other side, Deep Learning (DL) models are inherently better suited for capturing temporal dynamics. Researchers have been achieving high precision in FER through years, by

applying CNN with spatial data and, for sequential data, several works commonly use the combination between CNN-RNN especially LSTM network, indicating that CNN is the basic network of deep learning for FER, jointly with softmax function and Adam optimization algorithm [150]. Figure 2.9 shows an example of FER flow with deep learning.

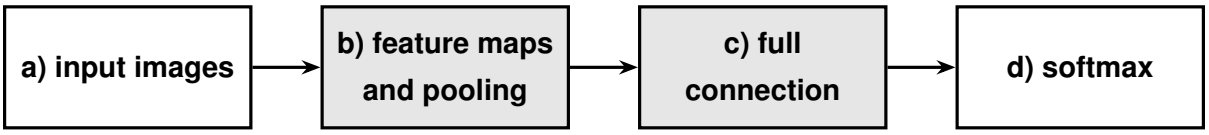


Figure 2.9 Typical FER flow with deep learning

Before 2000, there were few systems performing quantified facial expression classification into multiple basic emotion categories, which was only attainable with the development of computational resources [126]. Support Vector Machines (SVM), which dominated FER before deep-learning due to kernel-based nonlinear separability [124]. Dynamic Bayesian Networks model temporal dependencies in video sequences, capturing emotion dynamics [151].
CNN revolutionized FER, with architectures like AlexNet and ResNet achieving state-of-the-art results. Attention mechanisms further improve performance by focusing on salient regions like eyes for fear and mouth for joy. Transfer learning adapts pretrained models like VGG16 to FER, though dataset bias remains a challenge. Ensemble methods combining multiple CNN [152] and genetic algorithm-optimized architectures push accuracy boundaries on benchmarks like FER2013 (71.16%) and CK+ (98%) [124]. Figure 2.10 presents the timeline of FER solutions.

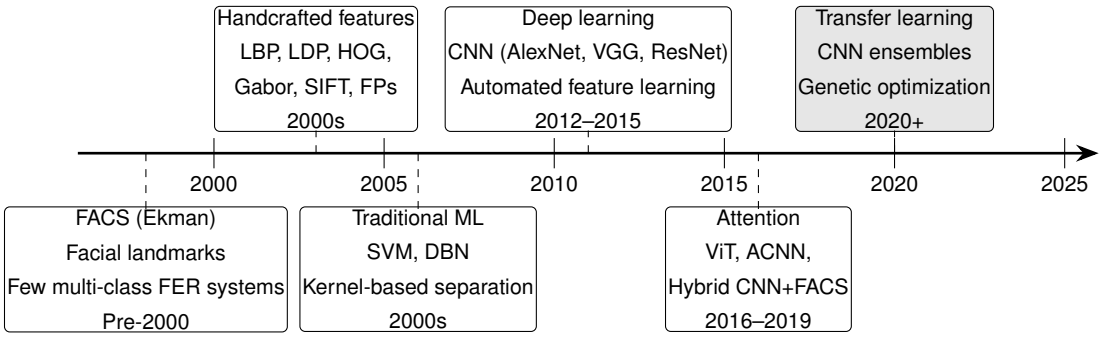


Figure 2.10 Timeline of facial emotion recognition (FER) development from handcrafted features to modern deep learning with attention mechanisms.

According to the survey by Bettadapura the error analysis from surveyed papers show that anger is confused with disgust, fear and sadness, while happiness and surprise are shown to be easier to classify [4].
During COVID19 period, Bo et al. produced a dataset of face images with artificial facial mask, to tackle the problem of FER with the mask occlusion [153]. Mushfieldt et al. tackles the presence of rotation and occlusion [154].

More recently, Konuk and Kiliç employed transfer learning with EfficientNetv2 architecture for robust feature extraction, leveraging attention mechanism trained with approximately 23.8 million parameters, classifying with 82.47% accuracy rate on the FER-2013 dataset [155].
Gursesli et al. proposed a lightweight CNN for facial emotion recognition based on MobileNetV2 architecture trained in AffectNet, RAF-DB and FER-2013 datasets, achieving comparable results on CK+, classifying seven emotions (92% of accuracy), 63% on FER2013, 84% on RAF-DB and 54% on AffectNet [7].
The DFEW dataset is highly imbalanced. For this reason, it is crucial to report both Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). WAR shows the overall performance, while UAR reveals whether the model is genuinely good at recognizing all emotions, or just the most frequent ones. A significant difference between WAR and UAR on a dataset like DFEW often indicates that the model struggles with minority classes.
We have found the following works with GitHub repositories:

- In 2023, Zheng et al. proposed **POSTER** [156], a solution that achieved state-of-the-art performance with accuracies of 92.05% on RAF-DB, and 67.31% on AffectNet. `https://github.com/zczcwh/POSTER`
- Recently, Sun et al. proposed **FCCA** (fast center consistency attention) [157], that achieved comparable performance while using less computational resources, with accuracies of 91.30% on RAF-DB, 65.51% on AffectNet, along with 56.61% UAR and 69.66% WAR on the DFEW dataset. The core idea is to perform convolutions on only a portion of the input channels, which significantly reduces the computational cost and memory usage without a major drop in performance. Feature extraction happens twice, once for the original image and again for a flipped version. This is a common data augmentation strategy to make the model more robust. `https://github.com/RuiSunooo/FCCA`
- This project implements real-time facial emotion recognition using the *DeepFace* library in combination with *OpenCV*. It captures live video from the webcam, detects faces in each frame, and predicts the corresponding emotional state. The identified emotion labels are then displayed on the video stream in real time. `https://github.com/manish-9245/Facial-Emotion-Recognition-using-OpenCV-and-Deepface/tree/main`

2.4 Multimodal Emotion Recognition (MER)

During a conversation, humans not only process the content of what is said, but also intonation, loudness, pauses, facial expressions, and even physiological signals. These elements

modulate the final meaning of a sentence and can even reverse its interpretation (as in cases of sarcasm or irony) [158].

MER arises from the need to enhance the robustness of a system by providing redundancy. If one modality is unavailable, the system can still function using the remaining data. Furthermore, latest advances in the field of emotion recognition have consistently shown that multimodal approaches outperform unimodal approaches in terms of accuracy and robustness, by fusing the diverse information present in different modalities. For example, the combination of vocal and visual cues allows the system to capture a more nuanced and complete representation of emotional expression.

MER task generally involves input data, which can be in the format of text, audio, or video, from which informative features are extracted (according to the data format) in a way that allows the prediction of the expressed emotion on the given input. The development of such systems rests on three fundamental methodological pillars: representation, alignment, and fusion. How each modality is represented, how they are synchronized in time, and how they are ultimately combined, directly determines the system's ability to capture and interpret emotional nuances. In this section, these challenges are addressed in turn: representation in Section 2.4.1, alignment in Section 2.4.2, and fusion strategies in Section 2.4.3. Finally, Section 2.4.4 reviews concrete methods and architectures for MER, illustrating how these principles are implemented in practice, from traditional statistical approaches to recent deep learning and Transformer-based solutions.

2.4.1 Multimodal Representations

In this section, the terms "features" and "representation" will be used interchangeably more frequently. The quality of the representations has a great impact on the performance of the trained machine learning models. According to [159], in multimodal problems, the distances between representations should reflect the similarity of the corresponding concepts, allowing to obtain missing modalities, based on the existing ones.

According to Baltrusaitis et al. [86], multimodal representations can be organized in two main categories:

1. **Joint representations**, correspond to early fusion approaches, where unimodal signals are represented in the same vector space.

2. **Coordinated representations**, in which the signals of each modality are processed separately, but projected into coordinate spaces, with enforced similarity constraints that allow correspondence between modalities.

2.4.2 Multimodal Alignment

As reviewed in Section 2.3, the extraction of unimodal representations has been widely studied, with the techniques shifting from hand-crafted features to deep features increasingly. In MER, the approaches typically involve splitting the input into utterances. An utterance can be defined as a unit of speech or text, which is bound by breathes, pauses, or end punctuations in the case of text [85].
Multimodal alignment is the process of finding correspondences between the representations of different modalities [86]. Three categories of multimodal alignment can be specified:

1. **Implicit alignment**, when representations of each modality have correspondences in a latent or intermediary multimodal space.

2. **Explicit alignment**, when the raw parts of both modalities are split based, for example, on the same window intervals. In these methods, features are extracted and can be immediately associated with the correspondent features of other modalities.

3. **Unsupervised alignment** is an approach in which correspondences are searched without resorting to explicit alignment labels, exploring statistical regularities between modalities.

2.4.3 Multimodal Fusion Techniques

In the context of MER, multimodal fusion refers to the process of combining information from different modalities with the aim of improving the performance of MER systems. The choice of fusion strategy influences not only the accuracy of the prediction, but also the robustness of the model in real-world scenarios, where not all modalities are always available.
There are two main categories of fusion: model-agnostic fusion and model-based fusion.
Model-agnostic fusion is preferred by a vast majority for multimodal fusion techniques due to its flexibility. Each fusion strategy differs from each other by when the fusion of different modalities actually occurs in the model pipeline.
Traditionally, three main categories are distinguished [86]:

- **Early fusion** consists of directly combining the features extracted from each modality, building a joint representation before the classification process, often by simply concatenating them. It can be **data-level fusion** or **feature-level fusion**. Data-level fusion can be applied when the raw inputs have the same temporal resolution and can be split in a way that we can make a correspondence with each part of each modality, while feature-

level fusion refers to combining different modalities into a single feature vector for emotion classification.

- **Late fusion**, or **decision-level fusion**, refers to processing each modality individually to train a classifier for each one and to use the output values of each model to make a decision, resembling ensemble models (multiple classifier systems). This approach uses unimodal decision values and fuses them with mechanisms such as averaging, voting schemes (example, majority voting), logistic regression, or weighting based on channel noise and signal variance [86]. The advantage of this approach is to offer greater modularity and resilience to the absence of modalities, allowing each model to focus on a single modality. A disadvantage is that the interaction between modalities is less explored, ignoring complementary information each modality might offer.

- **Hybrid fusion** combines the two previous paradigms, allowing for the integration of joint representations and independent decisions. This strategy can increase robustness without losing the richness of multimodal correlations.

Model-based fusion became more frequent with the advent of deep learning. Researchers started to apply model-level fusion, where features are learned jointly with classifier training. Deep fusion architectures using multimodal networks or attention mechanisms have shown state-of-the-art results.

Cross-modal attention networks allow one modality to condition the relevance of another modality's representations, reinforcing informative parts and attenuating redundancies.

Hierarchical models perform progressive fusion at different levels of abstraction, approximating the way humans integrate multiple signals.

Graph-based models, like GCN, represent modalities as nodes in a graph, explicitly exploiting the dependency relationships between them.

Multimodal transformers, like visual LLM, unify the processing of different modalities in a shared attention space.

Despite advances, significant challenges remain, mainly regarding temporal synchronization, data heterogeneity and missing modalities. Modalities such as audio and video may be misaligned, hindering early fusion. Different scales, dimensions, and noise can compromise joint representations. In real-world scenarios, not all modalities are always available, requiring robust imputation or inference methods. The selection of the fusion technique must balance accuracy, robustness, and computational efficiency, varying according to the application context.

The challenges of representation and alignment, combined with different multimodal fusion

strategies, establish the methodological basis on which MER systems are built. The way in which modalities are represented, synchronized, and subsequently combined directly impacts the model's ability to capture emotional nuances.

2.4.4 Classification Methods

In this Section, specific methods and architectures for multimodal emotion recognition (MER) are discussed, highlighting how fusion techniques are operationalized in different approaches, from classical systems to recent solutions based on deep learning and Transformers.
Over the years, many traditional and deep learning methods were explored by researchers, but for feature-level fusion, feature learning is vastly applied using deep learning approaches.

Audio-Visual

Before 2010, GMM was frequently used to classify emotion from audio-visual cues, fused at feature-level [160]. In 2010, Wöllmer et al. [161] proposed a context-sensitive MER solution, based on early (feature-level) fusion of acoustic and visual cues, using BiLSTM networks. This approach exploits long-range contextual information to model the evolution of emotion within a conversation, which was less explored until then. It focused on recognizing dimensional emotional labels (valence and activation values), fusing 30 from 46 selected (through PCA) face markers and a set of approximately 40 acoustic features.
In 2013, Kim et al. [162] presented a suite of Deep Belief Networks (DBN) models to study audio-visual feature learning in the emotion domain, through comparing unsupervised feature learning, using two-layer DBN enforcing multimodal learning, to secondary feature selection, before and after training. The accuracy maximum accuracy obtained was of 73.78% on IEMOCAP dataset.
With the advent of Transformers, more complex architectures were applied for MER. M3ER (Multimodal Memory Fusion Transformer) [163], proposed in 2020, introduces memory fusion to handle missing modalities and demonstrated robustness when one input modality is noisy or unavailable, reaching 82.7% mean accuracy on IEMOCAP.
RAVDESS, SAVEE, and CREMA-D are more recent datasets that are frequently used to evaluate MER systems. The feature-level fusion of facial and audio embeddings classified by LSTM [164] showed state-of-the-art accuracy of 88.11%, 86.75% and 80.27%, respectively.

Text-Audio

Studies focusing only in text and audio information are frequent, but typically obtain smaller accuracy than audio-visual solutions. Singh et al. [165] proposed unimodal and multimodal emotion recognition using a combination of 33 features (prosodic, spectral, and voice quality-based audio features), and additional textual embeddings from ELMo v2 (word and character embeddings), capturing context-dependent aspects of emotion in text, achieving 74.5% on IEMOCAP. which helped to capture the context-dependent aspects of emotion in text.
Recently, a solution of implicitly aligned multimodal transformer fusion (IA-MMTF) [166], based on acoustic features and text information was proposed, achieving 71.96% of accuracy.
SAVEE and RAVDESS are other datasets vastly used in literature to assess performance of SER and MER. In particular, [167] achieved 59.7% on SAVEE, using only acoustic features.

The following works with Github repository are related to this topic:

- **SpeechCueLLM (2024)** [168] proposed by Wu et al. propose a large language model framework that incorporates speech descriptions (seven long-term features) as auxiliary cues for emotion recognition. Their experiments on the IEMOCAP dataset show notable improvements: over 3 points in F1 score under the zero-shot setting (from 45.19% to 48.3%) and over 2.5 points under the LoRA fine-tuning setting (from 70.11% to 72.59%).
  `https://github.com/zehuiwu/SpeechCueLLM`

- **Emotion-LLaMA** [169], proposed by Cheng et al., is a fine-tuned LLM designed for emotion recognition tasks. The authors construct a multimodal dataset and leverage instruction tuning to enhance reasoning capabilities, enabling the model to capture complex multimodal emotional cues.
  `https://github.com/ZebangCheng/Emotion-LLaMA/`

- **DialogueLLM (2024)** [170] context and emotion knowledge tuned large language models for emotion recognition in conversations. This work integrates both textual and visual modalities, enriching conversational LLM with multimodal awareness.
  `https://github.com/X-PLUG/DialogueLLM`

- **Multi-Modality Collaborative Learning (MMCL)** [171], a framework for sentiment analysis that exploits cross-modal collaboration. It progressively decouples modality-specific representations and then refines them, capturing complementary emotional cues across modalities. Experiments demonstrate strong performance in multimodal sentiment pre-

diction.

https://github.com/smwanghhh/MMCL/

- **Cross-Modal Temporal Erasing Network (CTEN)** is a weakly supervised approach for video emotion detection and prediction. By strategically erasing temporal regions across modalities, the model avoids overfitting and learns more robust emotional features.

  https://github.com/nku-zhichengzhang/CTEN

- **Frame-Transformer Emotion Classification Network** operates at the frame level of videos. It models fine-grained temporal dependencies between frames for emotion classification, improving robustness in dynamic video settings.

  https://github.com/kittenish/Frame-Transformer-Network

- **MTCAE-DFER** [172] is a multi-task contrastive autoencoder for dynamic facial expression recognition, designed to leverage both reconstruction and contrastive learning objectives for robust emotion representation.

  https://github.com/Peihao-Xiang/MTCAE-DFER

- **MultiMAE-DFER** [173] is an adaptation of MultiMAE for dynamic facial expression recognition, incorporating masked autoencoding across modalities to enhance generalization in emotion classification.

  https://github.com/Peihao-Xiang/MultiMAE-DFER

- **Emotions (PyTorch implementation)** [174] is a repository containing implementations of neural models for facial emotion recognition, targeting both static and dynamic datasets.

  https://github.com/xuecwu/emotions

- **FV2ES (From Video to Emotion Signals)** [175] is a method that maps video inputs to continuous emotion signals, bridging discrete classification and dimensional affect prediction.

  https://github.com/qlwei89/fv2es

- **WECL (Weakly Supervised Cross-modal Learning for Emotion Recognition)** [8] is a framework that uses weak supervision to align modalities, improving robustness when full annotations are unavailable.

  https://github.com/nku-zhichengzhang/wecl

- **MIMAMO-Net** [176] model is focused on micro and macro expression recognition. It integrates hierarchical temporal modeling to handle subtle micro-expressions.

  https://github.com/wtomin/MIMAMO-Net, https://github.com/HKUST-NISL/MIMAMO-Net

- **Multimodal Transformer (MulT)** was proposed in 2019 and uses cross-modal attention to align audio, visual and text information for sentiment analysis and emotion recognition. This repository provides the original implementation, which inspired later adaptations such as MulT-Emo.

  `https://github.com/yaohungt/Multimodal-Transformer`

Chapter 3

Experimental Evaluation

This chapter presents the experimental evaluation of the proposed emotion recognition system. We begin by describing the computational environment in Section 3.1. Then, we conduct unimodal experiments for text (Section 3.2.1), speech (Section 3.2.2), and facial emotion recognition (Section 3.2.3). Each modality includes exploratory data analysis, baseline models, fine-tuning, and cross-corpus evaluations. Section 3.3 then integrates modalities into multimodal experiments, covering within-dataset, cross-corpus, and multi-corpus settings, followed by error analysis.

3.1 Computational Environment

The experiments were conducted on two different hardware setups.

- **Setup A (Central Processing Units (CPU)-based)**: MacBook Pro equipped with an Intel Core i9 processor (2.9 GHz, 6 cores), 16 GB DDR4 RAM (2400 MHz), and integrated Intel UHD Graphics 630 (1536 MB). This environment was used for all text and audio experiments involving traditional machine learning models.
- **Setup B (Apple Silicon)**: MacBook Pro with an Apple M1 Pro chip (8-core CPU, 14-core CPU) and 16 GB unified memory. This configuration was employed for visual emotion recognition experiments (DeepFace, CNN-based architectures) and LLM-based evaluations (Gemma2, GLM4, Qwen).

Both setups ran macOS with Python 3.11 and standard libraries (*NumPy*, *pandas*, *scikit-learn*, *Tensorflow*, *PyTorch*, *librosa*, and *openSMILE*). CPU acceleration was only available in Setup B via the Apple Metal backend, while Setup A relied exclusively on CPU computation.

3.2 Unimodal Experiments

In this section we will train and evaluate each machine learning component individually in datasets used in each of the unimodal emotion recognition task. Based on those evaluations, we then make decisions about models and settings to use in the multimodal system.
One limitation of these experiments is the number of emotions to classify in each dataset, which makes it harder to directly compare our study to others, since we chose to classify Ekman's basic emotions (anger, disgust, sadness, happiness, surprise and fear), plus the neutral emotion, interpreted as no emotion, while many studies focus solely on anger, happiness, neutral, and sadness.

3.2.1 Textual Emotion Recognition

Exploratory Data Analysis

The exploratory data analysis can be consulted in `EDA-TER.ipynb`.
EmoryNLP dataset does not have the anger emotion present, while does not contemplate the neutral category. None of the datasets has the seven emotions we study in this work, so we also evaluate TER on multimodal datasets. Figure 3.1 compares the emotion distributions in two of the TER datasets.

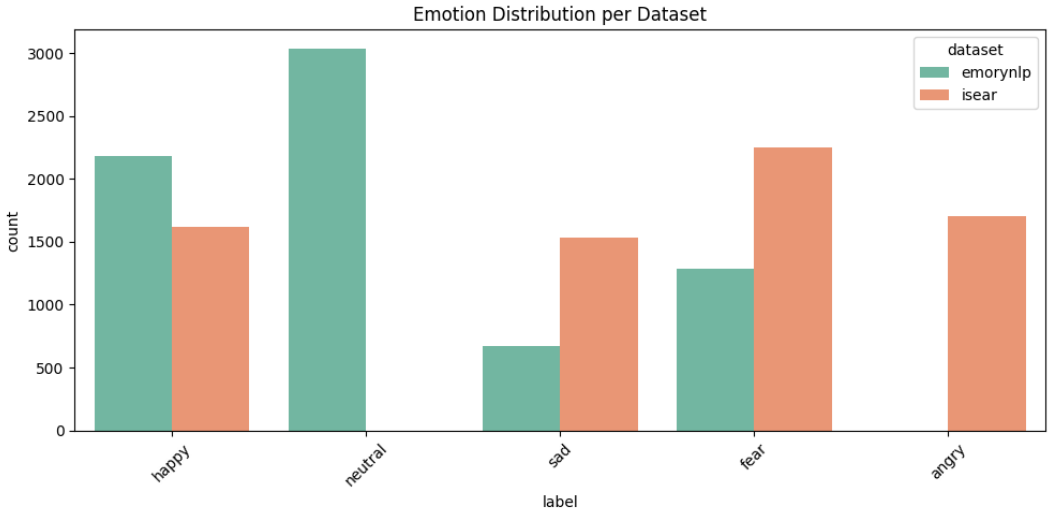


Figure 3.1 Countplots of filtered emotions of TER datasets (ISEAR and EmoryNLP).

When working with text data, the exploratory analysis usually focus on word count. Figure 3.2 shows the distributions of character length and word counts.
These datasets comprise short utterances, the majority below 40 words, with EmoryNLP having more than 2000 samples with less than five words, which is extremely short.

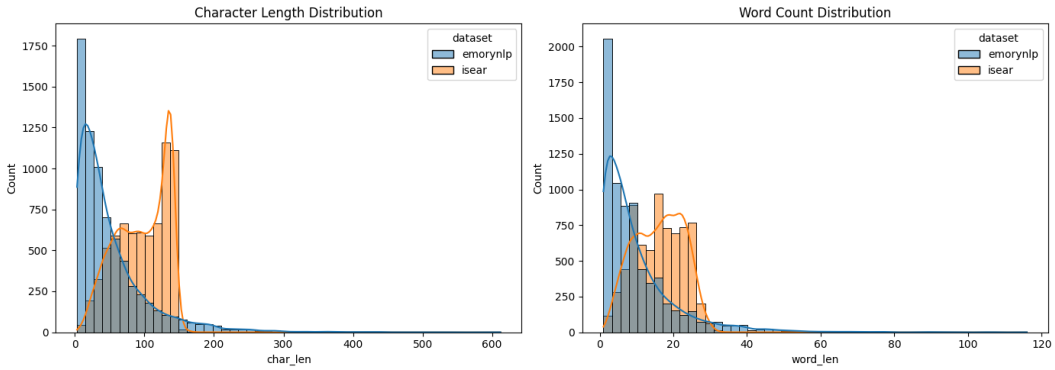


Figure 3.2 Distributions of character length and word counts.

Wordclouds are a informative way to look at the samples of each category and are shown in Figure 3.3.

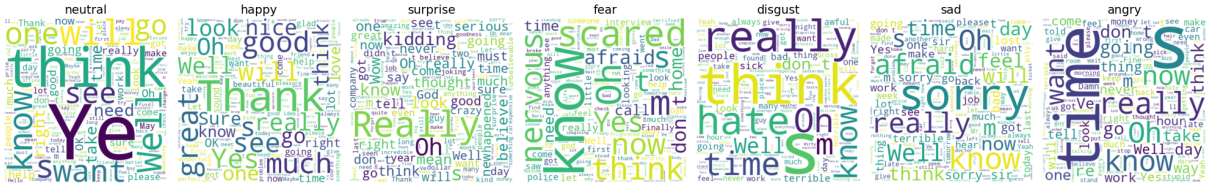


Figure 3.3 Wordclouds presenting the most frequent words in each emotion.

We can see that:

- The "happy" wordcloud shows the words "good", "great", "happy" and "love", which carry positive meanings.
- The "neutral" wordcloud shows the words "Ok", "Okay", "Hi" and "Hey", words of agreement and greeting, which normally do not convey any emotion.
- The "sad" wordcloud shows the words "sad", "sorry" and "lost", which convey sadness.
- In the "fear" wordcloud does not include any expressive words besides "fear", possibly making this emotion harder to classify.
- In the "angry" wordcloud shows the words "angry", "anger", "rage", "bitter" and "offended" which are very expressive words on this emotion.

Experiment 1: Evaluate Whisper Transcription

In this experiment, we evaluate the *base* and *tiny* Whisper models. Details on these results can be consulted at `02-exp1_2_3_4-TER-fine-tune.ipynb`.

Text From Audio Transcriptions: When applying TER to audio or video files, we must first extract the transcription of the audio of such files. We use Whisper, an Automatic Speech Recognition (ASR) model, to transcribe the audio signal. Whisper outputs time-aligned segments corresponding to individual phrases or sentences. We then slice the raw waveform according to these segment timestamps, producing audio chunks that align with meaningful linguistic units. These segments can subsequently be used for feature extraction. To evaluate Whisper, we will use Word Error Rate (WER), which is a standard metric in ASR. It measures the insertions, deletions and substitutions at the word level, given by:

$$\text{WER} = \frac{S + D + I}{N}. \tag{3.1}$$

As shown in Equation (3.1), WER is calculated using substitutions (S), deletions (D), and insertions (I) over the total number of words (N).

Table 3.1 presents the average of WER scores, and averages of insertions, deletions, and substitutions ratios for IEMOCAP and MELD. Both reference and transcription were lowercased and punctuation removed. We experimented with *base* and *tiny* models which have reasonable processing times on CPU systems, according to the Whisper documentation [177].

Table 3.1 Performance of Whisper *tiny* and *base* models on IEMOCAP and MELD.

Dataset	Model	Avg WER	Avg Insert.	Avg Delet.	Avg Subst.	Avg Hits
IEMOCAP	tiny	0.37	0.15	0.06	0.16	0.79
IEMOCAP	base	0.24	0.12	0.05	0.08	0.84
MELD	tiny	0.84	0.24	0.10	0.50	0.66
MELD	base	0.82	0.18	0.08	0.56	0.74

Results show models, namely the *base* model, perform significantly better on IEMOCAP, since it is an acted and relatively clean dataset, while MELD is composed by varied in-the-wild movie dialogues.

WER and Hits are not inverses of each other because WER includes insertions, but Hits does not penalize them directly. That is why, on MELD dataset, we simultaneously observe high average of WER and high average of Hits ratio, since we have a large number of correct words, but also a large number of insertions, which inflates WER.

Experiment 2: Train and Fine-Tune Models (Isolated)

In this set of experiments, we will train and evaluate strategies within TER datasets, such as EmoryNLP, ISEAR and DailyDialog, as well as MELD multimodal dataset, to verify results in

speech transcript. Details on these results can be consulted at `02-exp1_2_3_4-TER-fine-tune.ipynb`. Table 3.2 shows the results of training and testing approaches in the same isolated dataset.

Table 3.2 Performance of models on the textual modality. For traditional ML models, non-default hyperparameters from the *scikit-learn* implementation are indicated. Results correspond to the best configurations obtained through grid search.

Dataset	Model	Accuracy	Precision	Recall	F1-score
MELD	FastText + LR (C=0.1)	0.53	0.48	0.53	0.47
MELD	W2V + RF (n_est=100)	0.52	0.49	0.52	0.41
MELD	GLoVe + SVM (C=10)	0.46	0.38	0.46	0.38
MELD	DistilBERT (multilingual)	0.61	0.56	0.61	0.58
MELD	BERT(base)	0.62	0.60	0.62	0.60
ISEAR	DistilBERT (multilingual)	0.87	0.87	0.87	0.87
ISEAR	BERT (base)	0.87	0.87	0.87	0.87
DailyDialog	BERT (base)	0.84	0.78	0.84	0.79

Combinations were chosen based on 10-fold cross-validation grid search results, for each embedding method. Fine-tuning was performed using BERT and DistilBERT models for emotion recognition from text. Both models were initialized with pretrained weights available through the Hugging Face Transformers library, and the training process was carried out for three epochs, using the model-specific tokenizer with a maximum sequence length of 64 tokens. Training and validation were performed on CPU, with batch sizes of 4 and 8 for training and evaluation respectively. For optimization we used the AdamW optimizer with a learning rate of 5×10^{-5} and weight decay of 0.01.

Experiment 3: Train and Fine-Tune Models (Cross-Corpus Generalization)

The objective of this experiment is to evaluate the cross-corpus generalization capability of textual emotion recognition models, i.e, their ability to transfer knowledge learned from one dataset (source domain) to another unseen dataset (target domain). In this setup, models were trained on one dataset (Experiment 3.2.1) and tested on a different one without any additional fine-tuning on the target corpus. This evaluation helps assess the model's robustness and its dependence on corpus-specific linguistic or contextual biases. As shown in Table 3.3, results vary substantially depending on the source and target datasets. Among the evaluated corpora, MELD demonstrates relatively strong generalization performance across other datasets, achieving competitive accuracy and F1-scores regardless of the model architecture employed. This suggests that MELD's multimodal and context-rich conversational structure provides a

more diverse emotional representation that supports transferability. Details on these results can be consulted at `02-exp1_2_3_4-TER-fine-tune.ipynb`.

Table 3.3 Performance of cross-corpus generalization on the textual modality.

Train Data	Model	Test Data	Accuracy	Precision	Recall	F1-score
MELD	DistilBERT	DailyDialog	0.77	0.79	0.77	0.78
DailyDialog	BERT	EmoryNLP	0.43	0.38	0.43	0.31
DailyDialog	BERT	ISEAR	0.01	0.12	0.01	0.01
DailyDialog	BERT	MELD	0.50	0.33	0.50	0.36

In contrast, models trained on DailyDialog show limited generalization when tested on other corpora such as ISEAR and EmoryNLP. However, it is important to note that cross-corpus evaluation involving ISEAR and EmoryNLP was not fully conducted, as these datasets only include four emotion categories, which are not directly compatible with the seven or more emotion classes in MELD and DailyDialog, consequently, only compatible corpus pairs were evaluated.

Experiment 4: Train and Fine-Tune Models (Multi-Corpus Training)

Previous experiments focused on model performance within the confines of a single, source-specific dataset. However, real-world conversational systems require robust emotion recognition capabilities that generalize across diverse dialogue contexts, speaking styles, and annotation schemes. Table 3.4 shows the results of fine-tuning model DiltilBERT in more than one dataset. Details on these results can be consulted at `02-exp1_2_3_4-TER-fine-tune.ipynb`.

Table 3.4 Performance of multi-corpus training on text modality with DistilBERT.

Train Data	Test Data	Accuracy	Precision	Recall	F1-score
ISEAR; MELD; EmoryNLP	DailyDialog	0.49	0.70	0.47	0.57
EmoryNLP; MELD;	DailyDialog	0.81	0.74	0.81	0.77

Adding EmoryNLP to MELD dataset outputs higher accuracy than cross-corpus result (MELD –> DailyDialog), but with lower precision. This can be explained by both being dialogue-based datasets, which facilitates adaptation and boosts overall accuracy.

Experiment 5: Zero-Shot and Few-Shot Using LLM

Details on these results can be consulted at `03-exp5-TER_study-zero-shot-utt.ipynb`.

Tables 3.5 and 3.6 show performances of some LLM using zero-shot on MELD and IEMOCAP datasets, respectively.

Table 3.5 Zero-shot TER using LLM on MELD dataset (2610 valid examples).

Model	Nº Params	Accuracy	Precision	Recall	F1-score	Exec. Time
gemma	7 B	0.54	0.55	0.54	0.53	12m 45s
gemma2	9 B	0.55	0.60	0.55	0.57	16m 42s
glm4	9 B	0.54	0.58	0.54	0.55	11m 12s
qwen	4 B	0.53	0.47	0.53	0.47	7m 29s

On the MELD dataset, all models achieve moderate and consistent performance, with accuracies ranging from 0.50 to 0.55 and F1-scores between 0.47 and 0.57. The *gemma2* model obtains the best overall results, and the *glm4* model also performs competitively, while *qwen*, despite being the fastest model, shows the lowest F1-score, suggesting a trade-off between speed and performance.

Table 3.6 Zero-shot TER using LLM on IEMOCAP dataset (4639 valid examples).

Model	Nº Params	Accuracy	Precision	Recall	F1-score	Exec. Time
gemma	7 B	0.47	0.50	0.47	0.46	25m 28s
gemma2	9 B	0.48	0.55	0.48	0.50	32m 01s
glm4	9 B	0.45	0.51	0.45	0.46	20m 39s
qwen	4 B	0.45	0.51	0.45	0.39	14m 22s

In the IEMOCAP dataset, all models exhibit a clear degradation in performance, suggesting it is a more challenging dataset, but also that text modality is not sufficient to recognize emotions correctly. Overall, the zero-shot results indicate that all LLM capture general affective cues reasonably well but struggle with context-dependent emotions, that would be easier to identify through other modalities. The *gemma* family, particularly *gemma2*, appears to generalize more robustly across datasets.

Table 3.8 and Table 3.9 show performances using few-shot prompting on MELD and IEMOCAP datasets respectively.

Table 3.7 Zero-shot TER using *gemma2* and *glm4* on EmoryNLP (fear, happy, sad and neutral)

Model	Nº Params	Accuracy	Precision	Recall	F1-score	Exec. Time
gemma2	9 B	0.53	0.54	0.53	0.53	8m 2s
glm4	9 B	0.52	0.54	0.52	0.51	5m 28s
qwen	4 B	0.50	0.50	0.50	0.47	3m 35s

Table 3.8 and Table 3.9 show performances using few-shot prompting on MELD and IEMOCAP datasets respectively.

Table 3.8 Few-shot TER using LLM on MELD dataset (2610 valid examples).

Model	Nº Params	Accuracy	Precision	Recall	F1-score	Exec. Time
gemma	7 B	0.36	0.53	0.36	0.39	12m 46s
gemma2	9 B	0.48	0.64	0.48	0.50	16m 44s
glm4	9 B	0.54	0.59	0.54	0.54	11m 31s
qwen	4 B	0.50	0.47	0.50	0.43	7m 37s

When in-context examples are provided, model performance shows varied trends across datasets. On MELD, *glm4* achieves the best results overall, outperforming its zero-shot version slightly, but not zero-shot *gemma2*. Other models performed significantly worse, suggesting possible instability or over-sensitivity to few-shot prompting examples.

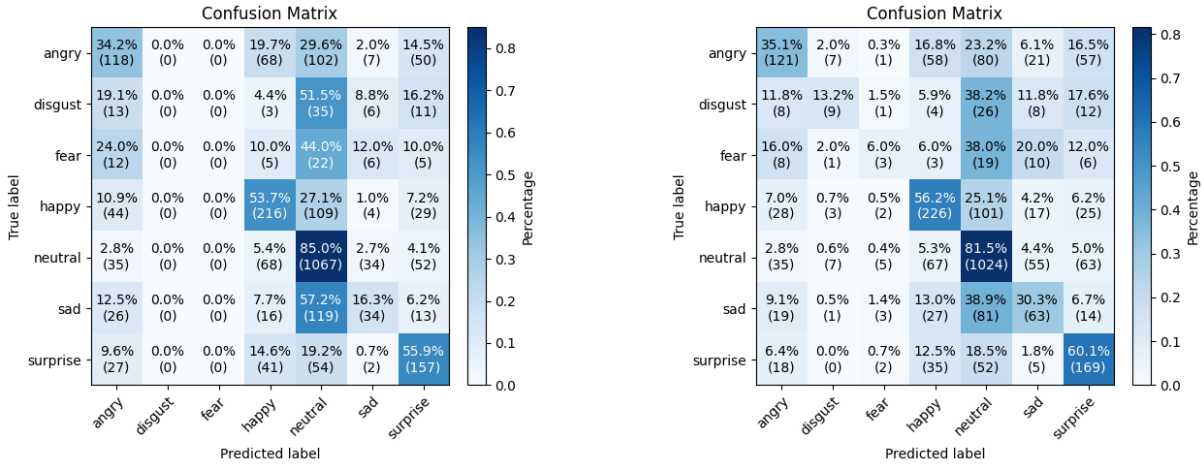
Table 3.9 Few-shot TER using LLM on IEMOCAP dataset (4639 valid examples).

Model	Nº Params	Accuracy	Precision	Recall	F1-score	Exec. Time
gemma	7 B	0.39	0.46	0.39	0.41	35m 10s
gemma2	9 B	0.43	0.56	0.43	0.42	42m 6s
glm4	9 B	0.45	0.52	0.45	0.46	33m 54s
qwen	4 B	0.43	0.52	0.43	0.35	18m 32s

In contrast, on IEMOCAP, the few-shot configuration does not consistently improve performance. The best model, *glm4*, which have 9 billion parameters, attains an F1-score of 0.46, identical to its zero-shot counterpart. All the other models degrade performance, indicating that few-shot prompting is not suitable for “small" LLM (< 7B parameters), which have limited context capacity, do not benefit from in-context examples. Execution times also increase across, as few-shot prompting requires longer contextual input per sample.

Error Analysis

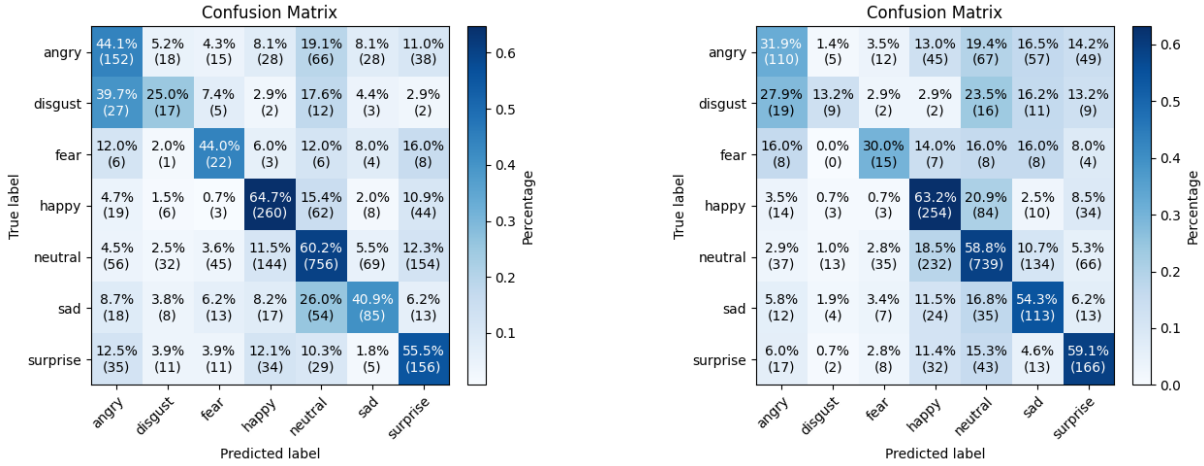
Confusion matrices allow to observe how misclassifications occurred. Figure 3.4 shows the confusion matrices of DistilBERT and BERT on MELD dataset. Multilingual DistilBERT does not correctly classify any of "disgust" or "fear" examples due to overfitting to "neutral" class, but BERT achieved the learning of those two classes and more true positives, although it also overfitted to "neutral".
Figure 3.5 shows the confusion matrices of *gemma2* and *glm4* on MELD dataset.



(a) DistilBERT model on MELD dataset.

(b) BERT model on MELD dataset.

Figure 3.4 Comparison of confusion matrices of DistilBERT (3.4a) and BERT (3.4b) on MELD dataset



(a) Gemma2 model on MELD dataset.

(b) Glm4 model on MELD dataset.

Figure 3.5 Comparison of confusion matrices of gemma2 (3.5a) and glm4 (3.5b) on MELD dataset.

Model *gemma2* performs best in every emotion except for "sad" and "surprise" emotions, and "disgust" is predicted as "angry" most of the times, indicating it did not learn well both emotions.

3.2.2 Speech Emotion Recognition

In this section we will train and test models both isolated and mixing more than one dataset, or testing on a test split from unseen dataset to verify the generalization of the model across different domains.

Exploratory Data Analysis

The exploratory data analysis can be consulted in `EDA-SER.ipynb`.

The chosen datasets are considered balanced in number of classes of each emotion. EmoDB and CREMA-D have only six emotions, because the "surprise" emotion is not present in those datasets, but it is present on RAVDESS and TESS. We filtered the "boredom" emotion from EmoDB, because no other dataset has samples of that emotion. Figure 3.6 shows the emotion distribution of some of SER datasets.

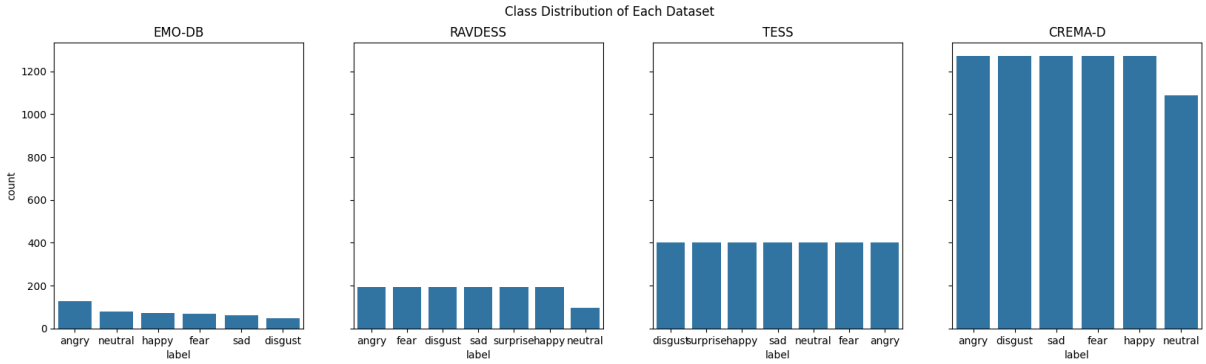


Figure 3.6 Countplots of filtered emotions of SER datasets.

Figure 3.7 shows the waveforms of one sample of each emotion, for each dataset.

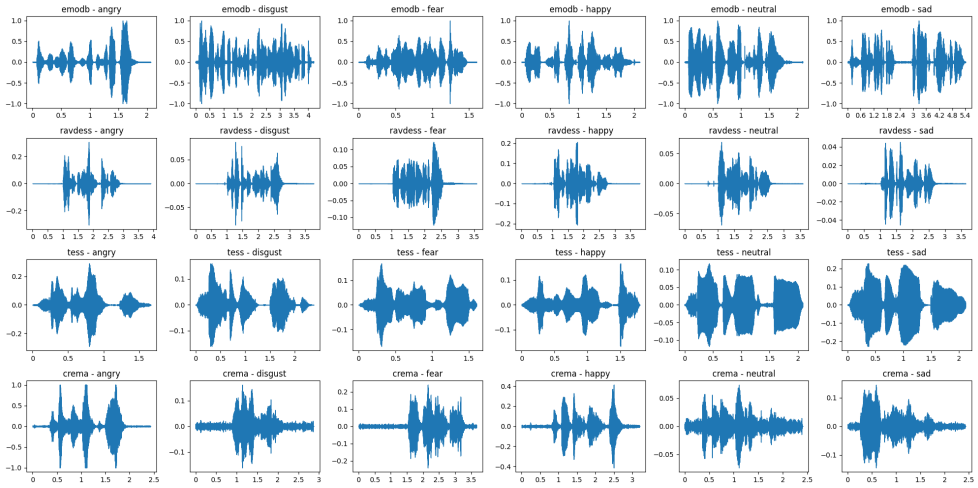


Figure 3.7 Waveforms of audios presenting different emotions, in the four chosen datasets.

We can observe that the waveforms are very characteristic from each dataset, indicating mismatching distributions and therefore challenging generalization.

Experiment 1: Traditional ML with and without feature selection

Details on these results can be consulted in files `02-exp1_2-SER_study_isolated.ipynb` and `02-exp1_2-SER_study_isolated_feature.ipynb`.

For a first experiment of SER we opted to test the GeMAPS feature set extracted with *openS-*

MILE, as it provides a compact representation that balances recognition accuracy with computational efficiency (62 features per sample). Given the lightweight goals of our system and the strong literature support for GeMAPS in SER tasks, emotional relevance and interpretability, we decided, in a second experiment to pursue additional feature selection, using an algorithm created by one of the authors. Table 3.10 shows some of the results obtained with and without feature selection.

Table 3.10 Traditional ML methods + GeMAPS features, with and without feature selection (isolated).

Dataset	Model	Accuracy	Precision	Recall	F1-score
EmoDB	SVM	0.81	0.83	0.81	0.82
EmoDB	FS + SVM	0.82	0.83	0.82	0.82
EmoDB	RF	0.76	0.77	0.76	0.75
EmoDB	FS + RF	0.78	0.78	0.78	0.77
EmoDB	LR	0.86	0.86	0.86	0.86
EmoDB	FS + LR	0.85	0.84	0.85	0.84
RAVDESS	RF	0.60	0.62	0.60	0.60
RAVDESS	FS + RF	0.52	0.54	0.52	0.52
TESS	RF	0.98	0.98	0.98	0.98
TESS	FS + RF	0.99	0.99	0.99	0.99
CREMA-D	SVM	0.52	0.52	0.52	0.52
CREMA-D	FS + SVM	0.53	0.53	0.53	0.53
CREMA-D	RF	0.52	0.51	0.52	0.51
CREMA-D	FS + RF	0.51	0.50	0.51	0.50

Overall, Logistic Regression achieved the best results, particularly on EmoDB. Feature selection showed slight improvements on TESS and CREMA-D (SVM) but degrading performance on RAVDESS. Models performed best on clean and balanced corpora (EmoDB, TESS) and worse on more heterogeneous datasets (RAVDESS, CREMA-D), indicating a strong dependence on dataset quality and emotional expressiveness.

Experiment 2: Train and Fine-Tune Models (Cross-Corpus Generalization)

Details on these results can be consulted in files `02-exp1_2-SER_study_isolated.ipynb` and `03-exp1_2-SER_study_isolated_feature.ipynb`.

Table 3.11 reports the cross-corpus generalization results for the audio modality. Models trained on the CREMA-D dataset were evaluated on unseen corpora (IEMOCAP and MELD) to assess their robustness to domain shifts.

The results indicate a clear drop in performance when testing on out-of-domain data. While

Table 3.11 Performance of cross-corpus generalization on audio modality.

Train Data	Model	Test Data	Accuracy	Precision	Recall	F1-score
CREMA-D	SVM (kernel=linear)	IEMOCAP	0.37	0.56	0.37	**0.43**
CREMA-D	RF + FS	IEMOCAP	0.38	0.55	0.38	0.42
CREMA-D	SVM (kernel=linear)	MELD	0.17	0.31	0.17	**0.19**
CREMA-D	RF + FS	MELD	0.13	0.26	0.13	0.11

the SVM with a linear kernel achieved a moderate F1-score of 0.43 when trained on CREMA-D and tested on IEMOCAP, its performance dropped significantly to 0.19 on MELD. This suggests that models trained on relatively clean and homogeneous corpora like CREMA-D struggle to generalize to datasets with higher variability in recording conditions, speakers, and emotional expressions such as MELD. Feature selection (FS) provided a marginal improvement, confirming that careful feature reduction may enhance cross-domain robustness but is insufficient to overcome domain-level disparities.

Experiment 3: Train and Fine-Tune Models (Multi-Corpus Training)

Table 3.12 presents the results of multi-corpus training experiments using SVM models across different combinations of audio datasets. Overall, the performance remains modest, indicating limited generalization when models are trained on heterogeneous emotional speech corpora and evaluated on unseen datasets. Details on these results can be consulted in files `04-exp3-SER_study_cross.ipynb` and `05-exp3-SER_study_cross_feature.ipynb`.

Table 3.12 Performance of multi-corpus training on audio modality.

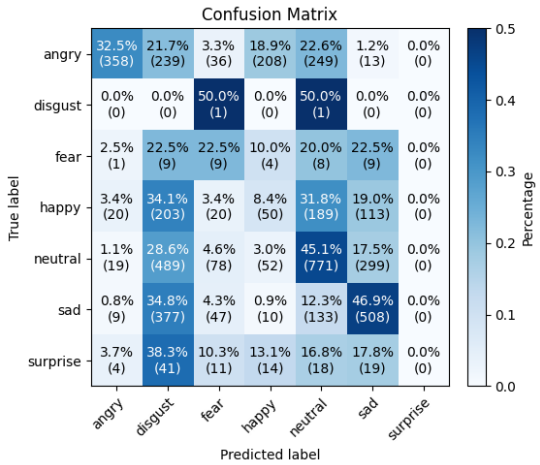
Train Data	Model	Test Data	Accuracy	Precision	Recall	F1-score
RAVDESS; TESS; MELD; IEMOCAP	SVM	EmoDB	0.48	0.40	0.48	0.33
RAVDESS; TESS	SVM(C=10)	IEMOCAP	0.14	0.46	0.14	0.13
RAVDESS; TESS; CREMA-D	SVM(C=10)	IEMOCAP	0.29	0.56	0.29	0.32

EmoDB, which is a German dataset, shows to be a challengeable domain to obtain generalization from the other studied datasets, that are spoken in English. IEMOCAP experiments including RAVDESS, TESS and ultimately CREMA-D, output worse performance metrics than cross-corpus generalization from CREMA-D alone, which misses surprise emotion (it does not

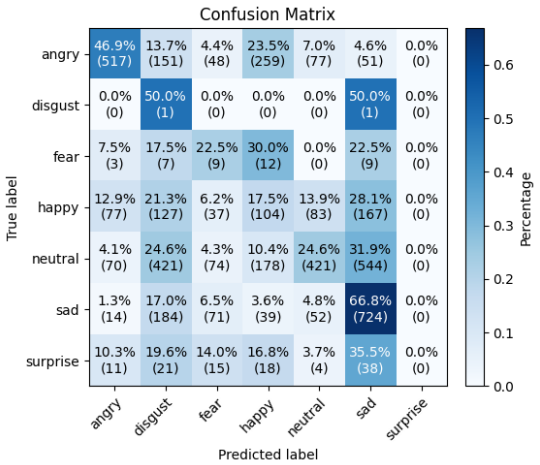
learn that class).

Error Analysis

As stated before, CREMA-D does not comprise the surprise label, therefore the model trained does not learn that emotion when trained only on CREMA-D. Figure 3.8 shows the confusion matrices of some multi-corpus experiments.



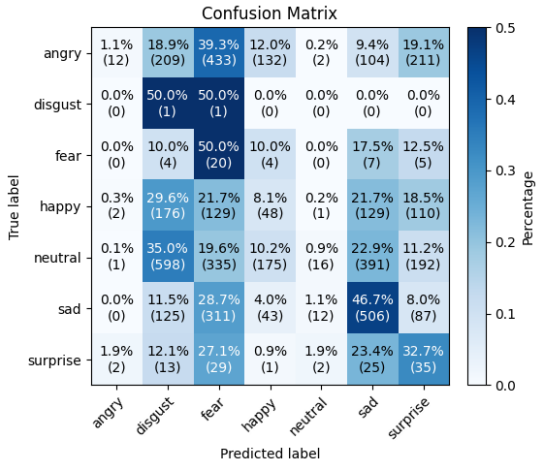
(a) CREMA-D Random Forest + Feature Selection model tested on IEMOCAP dataset.



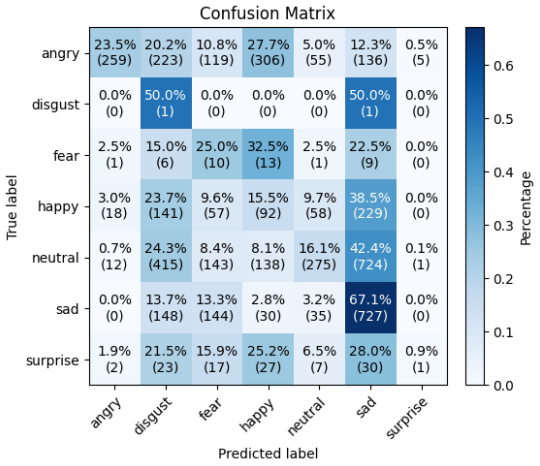
(b) CREMA-D SVM with linear kernel model tested on IEMOCAP dataset.

Figure 3.8 Comparison of confusion matrices of CREMA-D Random Forest + Feature Selection model (3.8a) and CREMA-D SVM with linear kernel (3.8b) on IEMOCAP dataset.

Figure 3.9 shows the confusion matrices of some multi-corpus experiments.



(a) RAVDESS + TESS model tested on IEMOCAP dataset.



(b) RAVDESS + TESS + CREMA-D model tested on IEMOCAP dataset.

Figure 3.9 Comparison of confusion matrices of RAVDESS + TESS model (3.9a) and RAVDESS + TESS + CREMA-D (3.9b) on IEMOCAP dataset.

The first matrix (3.8a), trained on CREMA-D and tested on IEMOCAP, shows many false positives on sad and neutral, and no true positives for emotion disgust. The surprise unpresent label is highly misclassified as disgust, and secondly, as sad and neutral. Happy emotion is highly misclassified as disgust and neutral. With linear kernel SVM (3.8b), we observe some improvements, disgust has one correct prediction (50% because IEMOCAP only has two samples) and all classes show improvement on true positives. Surprise emotion is frequently confused with sad, followed by disgust.
Using only RAVDESS and TESS (3.9a), we obtain high occurrence of disgust, fear, surprise and sad false positives. When adding CREMA-D (3.9b) there is a degradation of the performance of surprise emotion (since CREMA-D does not have this label), but improves or maintain the true positives of all other emotions, and also shows less false positives on disgust class.

3.2.3 Facial Emotion Recognition

Ablation tests with CK+ to test the best combinations of preprocessing techniques. Histogram equalization, Contrast Limited Adaptive Histogram Equalization (CLAHE), gamma correction, gaussian blur (bilateral filtering).

Exploratory Data Analysis

The exploratory data analysis can be consulted in `01-EDA-FER.ipynb`.
CK+ and FER2013 have their samples in grayscale, so we converted to RGB, which replicates the grayscale intensities through the three RGB channels. AffectNet and RAF-DB are BGR images, so they were converted to RGB as well. Figure 3.10 presents the emotion distributions of the four selected datasets.

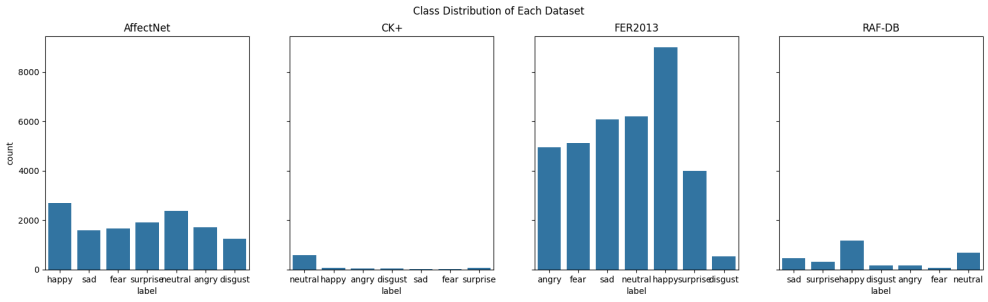


Figure 3.10 Countplots of filtered emotions of FER datasets.

By analyzing the countplots, we observe that AffectNet seems to be the most balanced dataset. CK+ has most of the samples from neutral emotion, FER2013 has very few examples of disgust, and RAF-DB has more of happy and neutral, than from the other emotions.

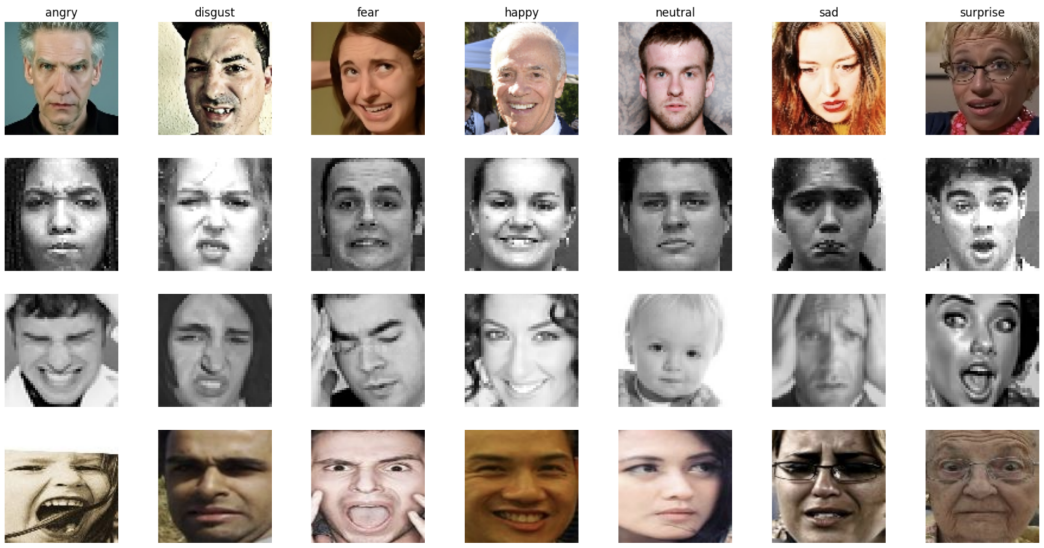


Figure 3.11 Examples of sample images for each emotion, for each dataset (AffectNet, CK+, FER2013 and RAF-DB).

All datasets consist of centered images. AffectNet and RAF-DB present RGB samples, while FER2013 and CK+ only include grayscale images.

To have some insight about the quality of the datasets and difficulty in recognizing emotions, we plot the average face of each emotion, for each dataset. AffectNet and FER2013 present less sharpened mean faces, which means the samples have higher variance than CK+ and RAF-DB. Figure 3.12 shows the average face of each emotion, for each dataset.

Figure 3.13 plots the Gabor filter response of mean face sharpened with highpass filter, for each emotion, for each dataset. We can observe that relevant edges can be extracted through this method.

Experiment 1: Deepface Library Different Backends

Details on these results can be consulted at `02-exp1-FER_study_results.ipynb`.

Image Processing: Images were cropped and resized to a consistent size, when training DL approaches. In the case of Deepface, image with minimal enhancement outputed best results, since preprocessing was worsing the performance. We can further augment the dataset with transformations like rotation, scaling, and flipping to increase variability.

On the CK+ dataset (Table 3.13), *retinaface*, *centerface*, and *skip* backends achieved the best performance, with *retinaface* offering slightly higher F1-scores but at the cost of considerably longer processing time (2m 26s for 93 samples). In contrast, the *skip* backend produced comparable results in only 1 second, demonstrating that when faces are already well-aligned or

Figure 3.12 Average faces for each emotion, for each dataset (AffectNet, CK+, FER2013 and RAF-DB).

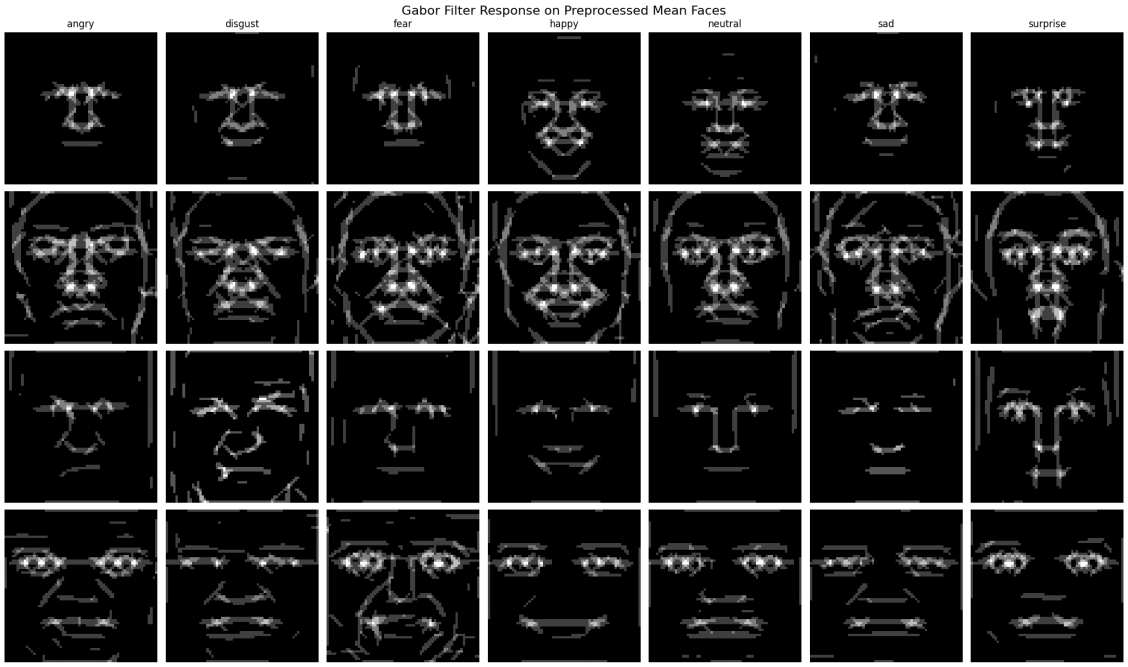


Figure 3.13 Gabor filter response of mean face sharpened with highpass filter, for each emotion, for each dataset (AffectNet, CK+, FER2013 and RAF-DB).

pre-cropped, skipping the detection step yields faster inference without compromising accuracy.

The performance achieved is not state-of-the-art (around 98%). This can be explained by the dataset size (which could be augmented) and the poor quality of images, presented in grayscale

Table 3.13 Performance of different DeepFace backends on the emotion recognition task, on CK+ test set, with 93 samples.

Model Backend	Total Time (93 Samples)	Accuracy	Macro F1	Weighted F1
opencv	2s	0.58	0.36	0.62
retinaface	2m 26s	**0.69**	**0.56**	**0.73**
ssd	3s	0.62	0.43	0.66
dlib	3s	0.65	0.39	0.65
ssd	3s	0.62	0.43	0.66
mediapipe	5s	0.62	0.42	0.66
yolov8	29s	0.66	0.49	0.71
centerface	12s	**0.69**	0.47	0.72
skip	**1s**	**0.69**	0.47	0.72

which conveys less information to a model that was trained on better resolution images.
For FER2013 (Table 3.14), a larger and noisier dataset, results were more balanced across detectors, with overall accuracies around 0.50–0.56. The *centerface* and *skip* backends again outperformed others, reaching a macro F1 of 0.55 while maintaining competitive inference times. These findings suggest that heavy detection backends like *retinaface* or *yolov8* offer no clear accuracy advantage for low-resolution or pre-aligned facial images, and that lightweight or skipped detection pipelines can provide a better trade-off between speed and accuracy in constrained or real-time applications.

Table 3.14 Performance of different DeepFace backends on the emotion recognition task, on FER2013 test set, with 3589 samples.

Model Backend	Total Time (3589 Samples)	Accuracy	Macro F1	Weighted F1
opencv	1m 7s	0.55	0.52	0.55
retinaface	1h -m -s	0.50	0.44	0.50
ssd	2m 56s	0.52	0.49	0.52
dlib	1m 24s	0.50	0.46	0.50
ssd	3m 35s	0.52	0.49	0.52
mediapipe	1m 1s	0.46	0.42	0.46
yolov8	15m 31s	0.51	0.47	0.51
centerface	7m 13s	0.56	**0.55**	0.56
skip	**42s**	0.56	0.54	0.56

Experiment 2: Train and Fine-Tune Models (Isolated)

Details on these results can be consulted at `03-exp2_3-FER_study_results.ipynb`.
We tested a lightweight CNN model that processes grayscale facial images of size 48×48, com-

prising four convolutional blocks with batch normalization and max-pooling, followed by a dense layer and a softmax output over seven emotion classes, as can be observed in Table 3.15.

Table 3.15 Architecture of the proposed lightweight CNN for emotion recognition.

Layer Type	**Filters / Units**	**Kernel**	**Activation**	**Obs.**	**Output Shape**
Input	–	–	–	Grayscale (48×48×1)	(48, 48, 1)
Conv2D + BN + ReLU	32	3×3	ReLU	Padding=same	(48, 48, 32)
MaxPooling2D	–	2×2	–	–	(24, 24, 32)
Conv2D + BN + ReLU	64	3×3	ReLU	Padding=same	(24, 24, 64)
MaxPooling2D	–	2×2	–	–	(12, 12, 64)
Conv2D + BN + ReLU	128	3×3	ReLU	Padding=same	(12, 12, 128)
MaxPooling2D	–	2×2	–	–	(6, 6, 128)
Conv2D + BN + ReLU	256	3×3	ReLU	Padding=same	(6, 6, 256)
MaxPooling2D	–	2×2	–	–	(3, 3, 256)
Flatten	–	–	–	–	(2304)
Dropout	–	–	–	rate=0.7	(2304)
Dense + ReLU	128	–	ReLU	–	(128)
Dense + Softmax	7	–	Softmax	Output layer	(7)

To improve the training efficiency and prevent overfitting, three Keras callbacks were employed during model training:

1. **ModelCheckpoint** — This callback monitors the validation accuracy after each epoch and saves the model weights corresponding to the best performance on the validation set, to ensure that the final model represents the best generalization state rather than the last training epoch.

2. **EarlyStopping** — Training is automatically stopped when the validation accuracy stops improving for a predefined number of epochs (patience = 20). This prevents unnecessary computation and helps avoid overfitting by restoring the weights from the best epoch.

3. **ReduceLROnPlateau** — When the validation loss plateaus for several epochs (patience = 3), the learning rate is reduced by a factor of 0.5, allows the optimizer to converge mwith stability toward the global minimum.

Table 3.16 presents the performance of training and testing the proposed lightweight CNN using a single dataset.

The proposed lightweight CNN demonstrates competitive performance when trained and tested on individual datasets. The model achieves its highest accuracy on CK+ (0.86), which can be attributed to the dataset's small size, controlled conditions, and limited variability. In contrast,

Table 3.16 Performance of proposed lightweight CNN on each isolated dataset (test=0.3).

Dataset	Accuracy	Precision	Recall	F1-score
FER2013	0.56	0.56	0.55	0.56
CK+	0.86	0.81	0.86	0.83
RAF-DB	0.66	0.68	0.66	0.65
AffectNet	0.60	0.60	0.60	0.60

performance on larger and more diverse datasets such as FER2013 (0.56) and AffectNet (0.60) is moderate, reflecting the increased difficulty caused by in-the-wild samples, variations in lighting, occlusions, and expression intensity.

Experiment 3: Train and Fine-Tune Models (Cross-Corpus Generalization)

Details on these results can be consulted at `03-exp2_3-FER_study_results.ipynb`.
Cross-corpus evaluations (Table 3.17) reveal a substantial performance gap when the model is trained and tested on different datasets. For instance, training on FER2013 and testing on CK+ results in an accuracy of 0.75, showing reasonable generalization when transitioning from a large, diverse dataset to a smaller, cleaner one. Conversely, the opposite setup drops sharply to 0.21, indicating poor generalization from controlled to in-the-wild domains.

Table 3.17 Performance of cross-corpus generalization on image modality with proposed CNN.

Train Data	Test Data	Accuracy	Precision	Recall	F1-score
FER2013	CK+	0.75	0.79	0.75	0.76
RAF-DB	CK+	0.46	0.63	0.46	0.50
AffectNet	CK+	0.51	0.60	0.51	0.54
CK+	FER2013	0.21	0.33	0.21	0.15
RAF-DB	FER2013	0.30	0.31	0.30	0.29
FER2013	RAF-DB	0.52	0.56	0.52	0.49
CK+	RAF-DB	0.23	0.36	0.23	0.11
FER2013	AffectNet	0.31	0.39	0.31	0.28
CK+	AffectNet	0.17	0.14	0.17	0.10
RAF-DB	AffectNet	0.18	0.20	0.18	0.16

These results highlight the dataset bias problem common in emotion recognition: models tend to overfit to dataset-specific lighting, demographic, and labeling distributions. Notably, training on FER2013 or AffectNet yields better cross-dataset performance due to their higher variability

and sample diversity. The limited generalization across datasets suggests that domain-specific characteristics strongly influence the learned representations.

Experiment 4: Train and Fine-Tune Models (Multi-Corpus Training)

Details on these results can be consulted at `04-exp4-FER_CNN_study_results.ipynb`.
The multi-corpus experiments in Table 3.18 demonstrate the benefits of combining datasets during training. Merging FER2013 and CK+ slightly improves performance on RAF-DB.

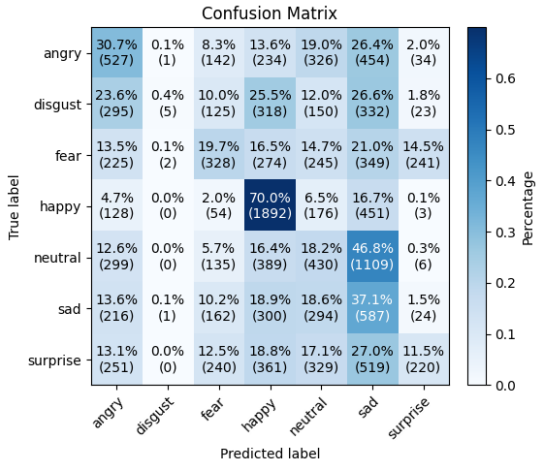
Table 3.18 Performance of multi-corpus training on image modality with proposed CNN.

Train Data	Test Data	Accuracy	Precision	Recall	F1-score
FER2013; CK+	RAF-DB	0.51	0.59	0.51	0.48
FER2013; CK+; AffectNet	RAF-DB	0.54	0.58	0.54	0.52
FER2013; CK+	AffectNet	0.30	0.34	0.30	0.27
FER2013; CK+; RAF-DB	AffectNet	0.34	0.43	0.34	0.31

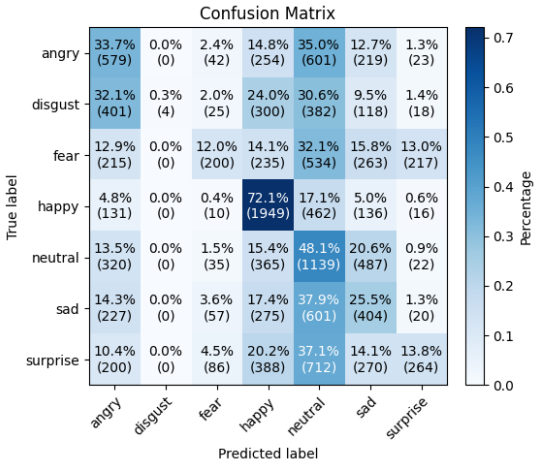
Notably, training on FER2013, CK+, and RAF-DB together significantly enhances AffectNet generalization, suggesting that the model benefits from exposure to a broader range of facial variations, illumination conditions, and emotional expressions.
This supports the hypothesis that data diversity is a key factor for cross-domain emotion recognition performance. Even a lightweight CNN benefits greatly from heterogeneous training sources, aligning with findings in recent literature that emphasize multi-corpus learning as an effective strategy to mitigate dataset bias.

Error Analysis

Figure 3.14 shows the confusion matrices of two multi-corpus experiments, tested on AffectNet. With FER2013 + CK+ configuration (3.14a), sad is the class presenting more false positives, suggesting subtle anger cues are often misinterpreted as neutral or positive emotions. Disgust remains challenging, with increased misclassification as sad, likely due to visual similarity among negative expressions.
Adding RAF-DB (3.14b) contributes greater diversity in facial expressions and lighting conditions, enhancing generalization. Misclassification patterns indicate that negative emotions (fear, disgust, sad) are often confused with one another or with neutral expressions, while positive emotions (happy, surprise) are generally well recognized. Low-frequency classes are particu-



(a) FER2013 + Ck+ model tested on AffectNet dataset.



(b) FER2013 + Ck+ + RAF-D model tested on AffectNet dataset.

Figure 3.14 Comparison of confusion matrices of FER2013 + Ck+ model (3.14a) and FER2013 + Ck+ + RAF-DB (3.14b) on AffectNet dataset.

larly prone to higher false positives, emphasizing the impact of class imbalance. These insights suggest avenues for improvement, including targeted data augmentation for underrepresented emotions.

3.3 Multimodal Emotion Recognition Experiments

This section explores multimodal fusion strategies for emotion recognition, integrating information from textual, acoustic, and visual modalities. Each modality was processed using specialized recognizers: a language model for text-based emotion recognition, a fine-tuned speech emotion classifier, and a facial expression recognizer applied to frame-level images. The outputs from these independent pipelines were combined through a fusion mechanism to derive a unified emotional prediction. Details on these results can be consulted at `1_late_fusion.ipynb`.

3.3.1 Fusion Experiments

This section explores the integration of multiple modalities to improve emotion recognition performance. The goal was to evaluate whether combining complementary sources of information could enhance robustness and overall classification accuracy compared to unimodal approaches.

There are three fusion strategies: *late fusion*, *early fusion*, and *hybrid fusion*. In **late fusion**, predictions from each modality are first obtained separately and then integrated through a decision-level fusion scheme. We implemented a flexible weighted fusion approach, capable

of handling both categorical labels (from traditional models such as SVM or logistic regression) and probabilistic outputs (from neural models like transformers or CNN). The model supports per-modality weighting and frame-level aggregation for the visual channel, which in the case of video file typically produces multiple emotion predictions, analyzing several extracted frames from the video.

Table 3.19 summarizes the results obtained from late, early, and hybrid fusion schemes across combinations of modalities. The abbreviations **T**, **A**, and **V** correspond respectively to the *textual*, *acoustic*, and *visual* modalities.

Table 3.19 Performance of multimodal emotion recognition models using late fusion. Weights of modalities are: T=0.4, A=0.3, V=0.3

Fusion Type	Modalities	Accuracy	Precision	Recall	F1-score
Late Fusion (with face detection)	T + A	0.45	0.47	0.45	0.46
Late Fusion (with face detection)	T + A + V	0.35	0.46	0.35	0.39
Late Fusion (without face detection)	T + A	0.45	0.47	0.45	0.46
Late Fusion (without face detection)	T + A + V	0.37	0.46	0.37	0.40

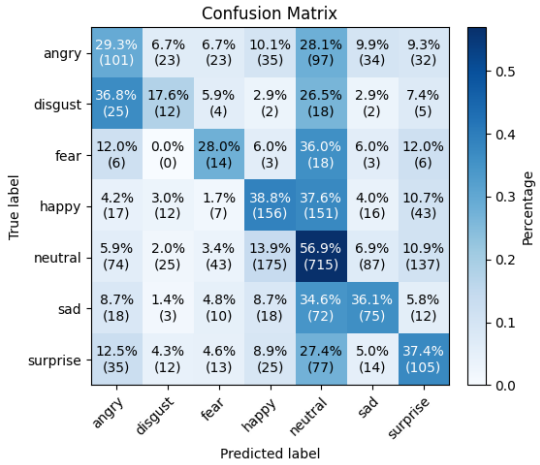
We tested fusion of results of the following components: text modality uses *gemma2* model with zero-shot prompting; audio modality uses the best SVM model trained on CREMA-D dataset, without feature selection, which outputted interesting results on MELD; visual modality is implemented as collecting one frame per second and identifying emotion in faces using *Deepface* "skip" backend.

When comparing these results with the cross-corpus SER experiments in Table 3.11 we observe the accuracy improves from 0.17 to 0.45, when fusing text and audio scores. But unimodal text results from Table 3.5, using zero-shot prompting is still higher than the performance obtained with fusion.

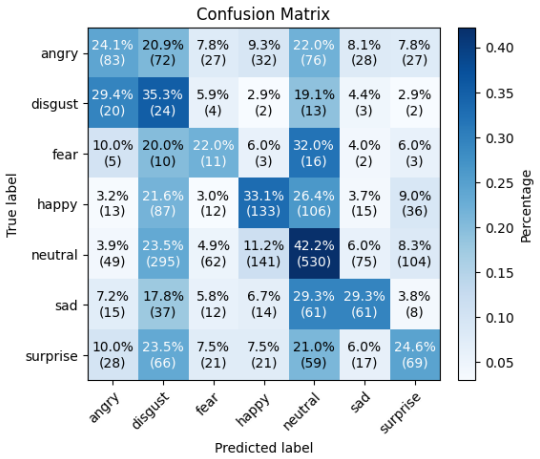
3.3.2 Error Analysis

Figure 3.15 shows the confusion matrices of two late fusion experiments using face detection as a preprocessing step, evaluated on the MELD dataset.

The addition of the visual modality led to a noticeable degradation in performance, likely due to the strategy employed for frame collection and modality scores fusion. Given the natural head movements of speakers in the video clips, sampling one frame per second often failed to capture proper frontal faces, resulting in misaligned or partially occluded visual inputs. Consequently, the visual stream introduced noise instead of complementary information, diminishing the overall effectiveness of the trimodal configuration. These findings suggest that the reliability



(a) Late fusion text + audio model tested on MELD dataset.



(b) Late fusion text + audio + visual model tested on MELD dataset.

Figure 3.15 Comparison of confusion matrices of text + audio model (3.15a) and text + audio + visual model (3.15b) on MELD dataset.

and temporal consistency of facial features are critical for successful multimodal integration. Due to time constraints, early and hybrid fusion strategies—potentially capable of capturing cross-modal dependencies more effectively—were not explored in this work. They are identified as promising directions for future research, particularly when combined with improved face tracking, adaptive frame selection, and synchronized multimodal representations.

Chapter 4

Implementation

This chapter presents the implementation of the proposed multimodal emotion recognition system and its corresponding application. Section 4.1 introduces the overall concept and user interaction flow of the application, detailing how multimedia inputs are processed and analyzed through emotion recognition and conversational feedback. Section 4.2 defines the lightweight and deployment constraints that guided this implementation. Section 4.3 outlines the modular architecture integrating text, speech, and facial emotion recognizers through a late fusion strategy, supported by a Python *FastAPI* backend and a React-based frontend. Finally, Section 4.4 describes the software design and implementation of these components, detailing their interactions, data flow, and modular organization.

4.1 Application Overview

The proposed system aims delivers a proof-of-concept system that integrates the machine learning components developed in this thesis. This application can be used to help students and workers prepare presentations, helping them on public speaking, but it can also be used to train specific answers to a list of probable questions of a job interview.

The application flow starts when a user uploads their multimedia content, that can be in the format of text, image, audio or video recording. Upon successful upload, the system transitions into the ER phase. This phase employs trained classifiers to examine visual or acoustic properties, and LLM to examine the semantic content of the text, with the objective of correctly classifying the expressed emotions within these modalities. The output of the ER phase is a series of emotion labels, each precisely time-stamped and associated with a confidence score, is then passed to the next phase.

In the chatbot module, an LLM-powered agent assumes the role of an emotional analyst. It undertakes a detailed examination of the ER output, identifying patterns and temporal distributions

of emotional expressions. This analysis generates a comprehensive report that enumerates the detected emotions and contextualizes them within the flow of the user's communication, revealing potential correlations and insights. This report is then presented to the user.

The user is then invited to engage in a natural language dialogue with the chatbot. This interactive phase transforms the static report into a dynamic exchange. Users can ask questions, seek clarifications, and delve deeper into the nuances of their emotional expressions. The chatbot, acts as a personalized emotional intelligence coach, providing details and actionable advice, helping the user analyzing his data. This flow allows users to explore the "why" behind their emotional patterns and to consider strategies for improvement. Figure 4.1 presents and overview of the proposed system's flow.

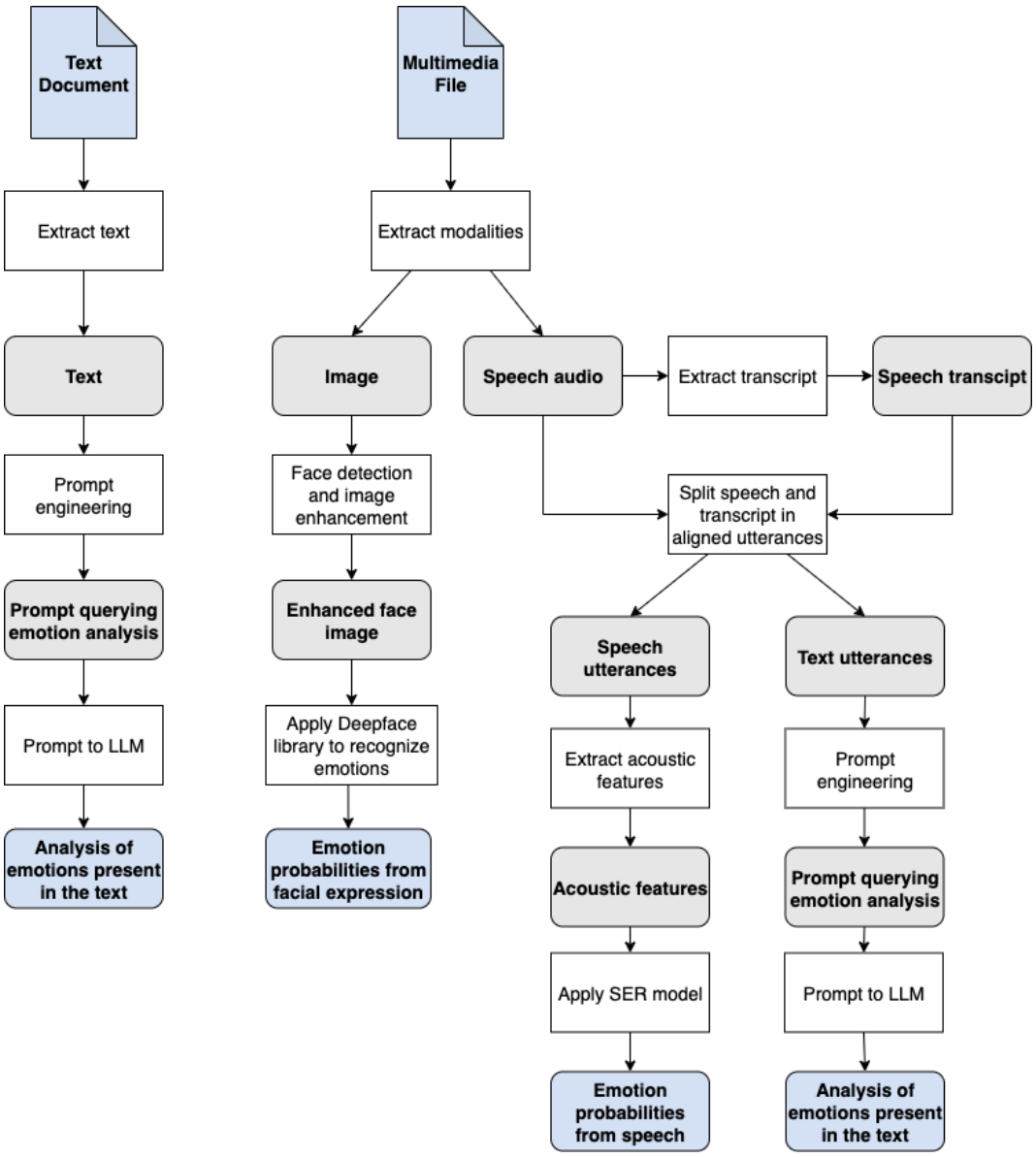


Figure 4.1 Multimodal emotion recognition application flow diagram.

4.2 Definition of Lightweight and Deployment Constraints

In this work, the term “lightweight” refers to models and system components that are optimized for efficient execution on standard CPU environments, enabling real-time or near real-time interaction for single-sample inference. Specifically, a lightweight model in this context satisfies the the criteria presented in Table 4.1.

Table 4.1 Lightweight system criteria.

Criterion	**Definition**	**Target / Limit**
Computational Efficiency	Maximum allowable inference time per single input sample on a standard CPU, ensuring real-time or near real-time responsiveness.	Text / Audio: <2 s; Video: <5 s
Memory Footprint	Total memory consumption during inference, measured to guarantee compatibility with consumer-grade hardware without Graphics Processing Units (GPU) acceleration.	<4 GB
Model Complexity	Scale of trainable parameters and model size to balance performance and efficiency. to reduce memory and computational requirements, while maintaining competitive performance. Includes LLM used for textual emotion recognition and conversational AI.	<7 billion parameters
Modularity and Deployment Flexibility	Ability for each modality-specific component (Text, Speech, Facial) to operate independently, allowing selective execution depending on scenario and resources.	Independent execution per modality

Inference for a single input sample (text, audio, or visual) completes within a reasonable timeframe on a standard CPU (<2 seconds for text or audio clips of typical length, and <5 seconds for videos). It is important to note that LLM are employed for textual emotion recognition to leverage their advanced contextual understanding. While full-scale training and fine-tuning of these models are computationally intensive, the proposed system focuses on CPU-based inference for individual inputs, aligning with the lightweight objective. This ensures that end-users can interact with the system in real-time, receiving personalized feedback without access to high-end hardware.

4.3 System Architecture

The proposed multimodal emotion recognition system follows a modular, end-to-end architecture that integrates three unimodal emotion recognition pipelines as shown in Figure 4.2.

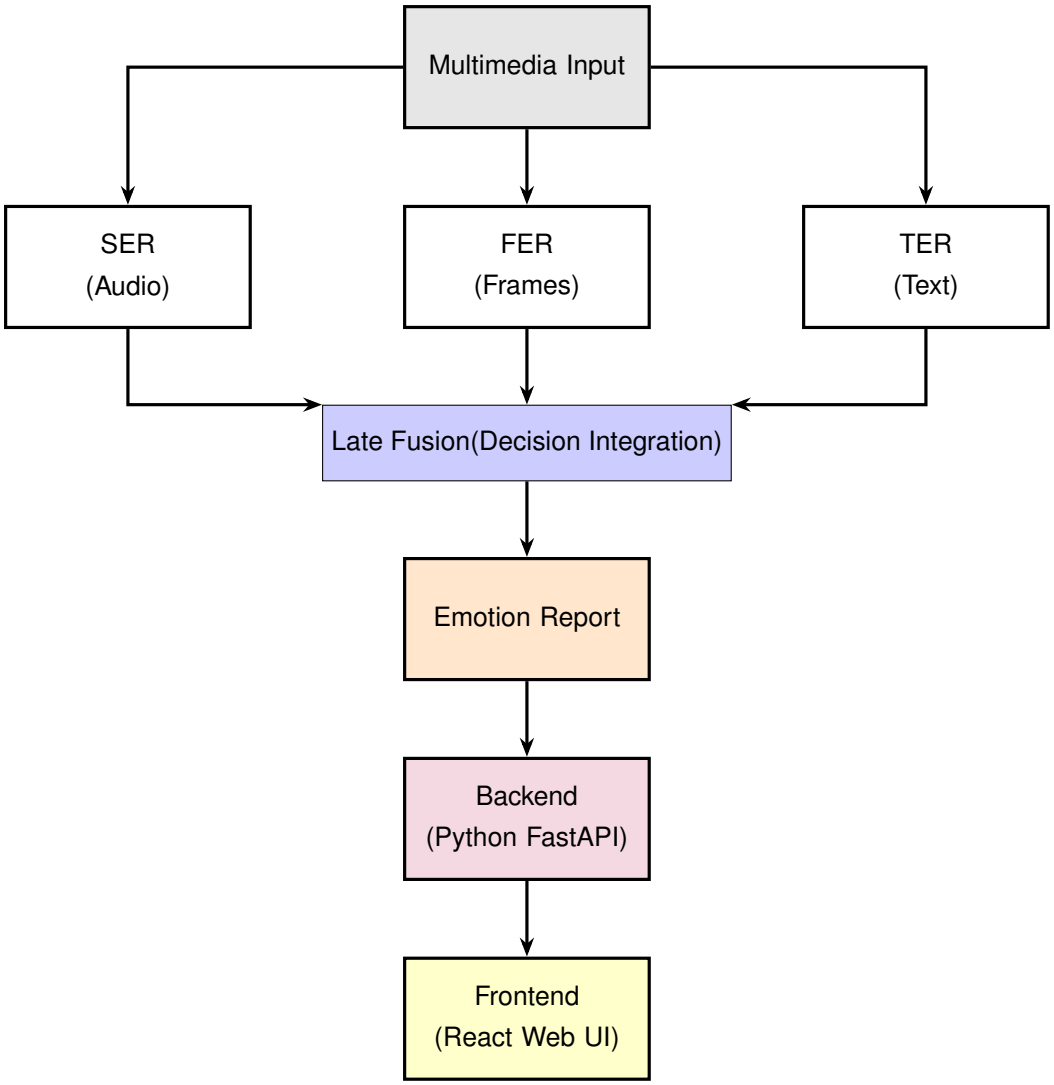


Figure 4.2 Overview of the proposed multimodal emotion recognition system, showing the three unimodal recognition modules integrated through late fusion, followed by backend and frontend components.

Multimedia inputs (audio, video frames, and text) are processed to extract modalities and classify them by their respective unimodal recognition modules. **SER** extracts acoustic features from audio signals and classifies emotional tone using a trained CNN or RNN model. **FER** analyzes facial frames using a lightweight CNN or *Deepface* to infer visual emotion cues. **TER** processes textual data through prompting or transformer-based classifier to detect linguistic sentiment.

The outputs of these three modalities are integrated in the **Late Fusion** stage, which combines predictions at either the feature or decision level to obtain a unified emotional representation. This fused result generates the final **Emotion Report**, which is handled by the **Backend** (implemented with Python *FastAPI*) for further processing and transmitted to the **Frontend** (developed in React) for visualization and user interaction.

4.4 Software Design

The implementation of this application is available as a GitHub repository. The project repository code is divided into two main sub-folders:

- **Fontend** repository (emorea-frontend): React code with respect to the client-side of the application, responsible for presenting information to the user through the user interface (UI). This project uses "npm" for dependency and package management.
- **Backend** repository (emorea-backend): Python code with respect to the server-side. Packages and dependencies are managed by "poetry" Python library.

Figure 4.3 maps the interactions between low-level components and modules from frontend and backend.

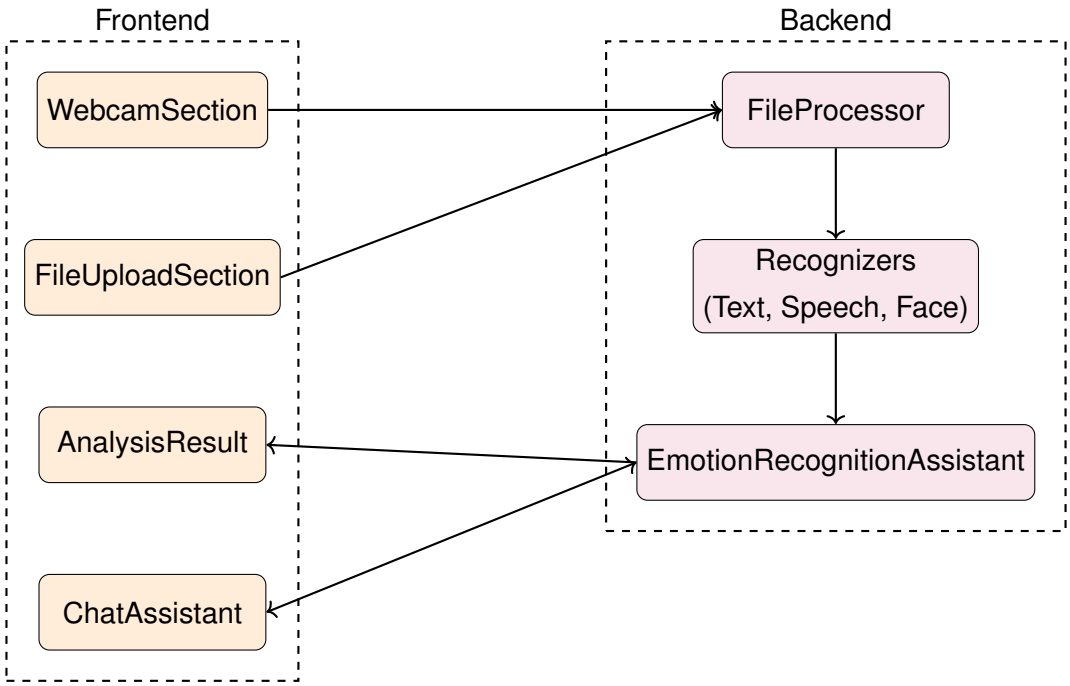


Figure 4.3 Architecture diagram: frontend components and backend modules with interactions.

The frontend components WebcamSection and FileUploadSection capture visual input and sends it to the backend FileProcessor for preprocessing. The processed data passes through the Recognizers module, which analyzes text, speech, and facial features. The EmotionRecognitionAssistant integrates these results to infer the user's emotional state, returning responses to the AnalysisResult and ChatAssistant interfaces for user feedback and interaction.

4.4.1 Backend Implementation

The backend of this application was written in Python, and FastAPI was used to build the server side. We adopted a modular architecture with clear separation of concerns:

FileProcessor (preprocessing and ingestion) identifies file format and extracts modalities present in the file, making them ready for feature extraction.

Recognizers (inference) are a group of classes that perform feature extraction from raw modalities and apply the corresponding classification model. The classes are:

- **TextEmotionRecognizer** leverages LLM-based classification for emotion classification, via litellm to an Ollama endpoint. Since LLM queries are inherently blocking, a ThreadPoolExecutor is employed to enable concurrent processing of multiple requests without stalling the main pipeline. Batch inference is supported, where multiple texts are submitted in parallel and results are returned in the same order as input. This design ensures both scalability and robustness in real-time applications. Minimal preprocessing is performed here, since the input is already natural language, thus only prompting strategy is applied.

- **SpeechEmotionRecognizer** loads a *joblib* pipeline (operating the chosen model) and computes acoustic features for inference. The default feature extraction method is using *openSMILE*, extracting GeMAPSv01b functionals, a robust paralinguistic set. Librosa-based descriptors are also implemented, extracting MFCC, chroma, and Mel-spectrogram features. This recognizer purposefully does not perform STT/transcription; that function is owned by FileProcessor.

- **FaceEmotionRecognizer** employs DeepFace library with a configurable detector backend (MTCNN, RetinaFace, OpenCV) to classify emotions from images or frames. Preprocessing beyond color conversion is intentionally minimal here; face detection/cropping for video is handled in FileProcessor. For robustness, the recognizer automatically handles grayscale-to-RGB conversion and disables strict face enforcement to avoid failure in low-quality frames. Outputs include both the full probability distribution across emotions and the dominant predicted label. For videos, analysis is performed frame-by-frame, with optional optimizations to balance speed and accuracy.

- **EmotionRecognitionAssistant (orchestration)** coordinates preprocessing, routes artifacts to recognizers, aggregates results, and (optionally) feeds the LLM with the analysis summary. This controller enforces modality independence, and allows late fusion or report generation in a single place.

Figure 4.4 maps the interactions inside the backend component.

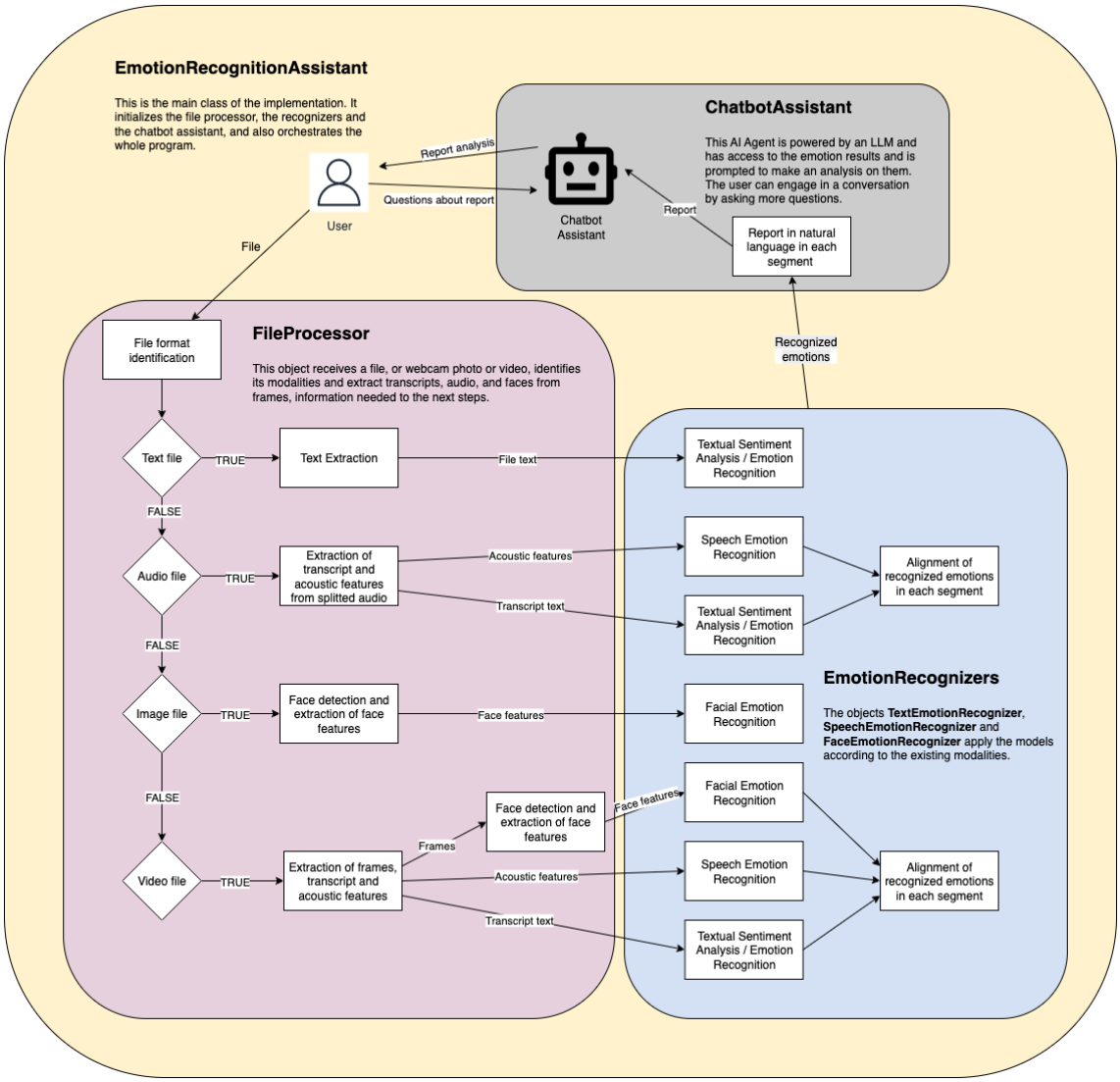


Figure 4.4 Diagram of interactions between the backend components.

This decomposition aligns with the Single Responsibility Principle and supports testability, substitution of components, and future extensions (alternative STT, facial backends, or fusion strategies), so this project is easy to extend and debug each component independently.

4.4.2 Frontend Implementation

The frontend of the application, illustrated in Figure 4.3, was implemented in React, following a modular component-based architecture. Each component encapsulates a distinct functionality, allowing for clear separation of concerns and easy scalability. The interaction between these components and the backend modules is represented by directional arrows in the diagram.

- The **WebcamSection** and **FileUploadSection** components handle user input acquisi-

tion. The former manages real-time video and image capture using the react-webcam and recordrtc libraries, while the latter provides a drag-and-drop interface for uploading multimedia files (audio, video, text, or image). Both components send their data to the backend's FileProcessor through HTTP POST requests, initiating the emotion recognition pipeline.

- The **AnalysisResult** component displays the model predictions returned from the backend. Results are presented in a structured JSON format, containing emotion labels, confidence scores, and metadata related to each analyzed segment, this structure can then be converted to natural language.

- Once the analysis phase concludes, the **ChatAssistant** component establishes an interactive channel between the user and the backend's EmotionRecognitionAssistant. This chatbot interface allows users to interpret the emotional analysis, request clarifications, or receive personalized feedback. Communication is achieved via RESTful endpoints exposed by the FastAPI server.

State management is handled locally via React hooks (useState, useRef), avoiding unnecessary dependencies.
Together, these components form the client-facing layer of the multimodal emotion recognition system, directly corresponding to the left (orange) section of Figure 4.3. Their communication with the backend's modules (FileProcessor, Recognizers, and EmotionRecognitionAssistant) ensures a seamless and interactive user experience.

Chapter 5

Conclusions and Future Work

This chapter summarizes the main findings of this research. Section 5.1 discusses the overall experimental insights and performance across modalities. Section 5.2 presents representative case studies that illustrate practical applications of the proposed system. Section 5.3 outlines the main challenges encountered during development and evaluation, and future work directions and possible system extensions.

5.1 Discussion and Insights

5.1.1 Text Emotion Recognition (TER)

Large Language Models (LLM) proved effective in inferring emotions directly from transcribed speech or dialogue text. However, not all LLM were equally capable: smaller models often deviated from instructions by outputting multiple emotions or morphologically related terms instead of adhering to the constrained emotion list.

One advantage of prompt-based LLM emotion recognition lies in its adaptability, since changing the emotion list prompted to the LLM requires no retraining, allowing seamless transfer across datasets with different label schemes. Despite this flexibility, textual emotion detection remains highly context-dependent and limited in cases where emotional tone is conveyed primarily through intonation or facial expressions.

5.1.2 Speech Emotion Recognition (SER)

The acoustic modality captured prosodic cues such as pitch, energy, and temporal dynamics that are often absent from textual data. Nevertheless, results highlighted significant speaker and recording variability, which affected model robustness, particularly in cross-corpus conditions. This suggests the need for domain adaptation through data augmentation to handle

differences in recording environments, microphones, and speech styles.

5.1.3 Facial Emotion Recognition (FER)

Experiments on four benchmark datasets (CK+, FER2013, RAF-DB, AffectNet) confirmed the influence of data quality, color information, and dataset bias on model performance.
DeepFace backends showed that heavy detectors (e.g., RetinaFace, YOLOv8) did not necessarily outperform lightweight alternatives when images were already well-aligned. The skip and centerface configurations achieved competitive results at a fraction of the computational cost, making them ideal for real-time deployment. Datasets for FER are massively available, but few are in RGB colors and with good quality, so Deepface which was trained in proprietary data and did not present the best performance in those datasets, since other state-of-the-art models are mostly trained in the stricted universe of those datasets. Overall, DeepFace's performance remains below state-of-the-art results reported for FER2013 (typically 65–70%), likely due to differences in preprocessing, lighting conditions, and the limited adaptation of pretrained models to grayscale inputs. Nonetheless, these experiments confirm the robustness of the framework and provide insights into optimizing backend choice for deployment scenarios balancing performance and computational efficiency.
The lightweight CNN proposed in this work achieved competitive performance (up to 0.86 accuracy on CK+), demonstrating the feasibility of emotion recognition using small, efficient models. Cross-corpus results, however, revealed poor generalization, confirming that models tend to overfit to dataset-specific characteristics such as illumination, demographics, or capture conditions. Multi-corpus training partially mitigated this issue, leading to improved robustness, especially when combining FER2013, CK+, and RAF-DB.

5.1.4 Multimodal Emotion Recognition (MER)

Due to time constraints, only the Late Fusion approach was fully implemented and evaluated experimentally. This strategy offered several practical advantages: it preserved the modularity of each unimodal recognizer, simplified the integration pipeline, and allowed straightforward calculation with modality emotion confidences. Moreover, since each recognizer outputs emotion probabilities independently, late fusion can also provide interpretability by showing the contribution of each modality to the final decision. More work needs to be done to improve this method.
Early and hybrid fusion are also future directions.

5.2 Representative Case Studies

The idea of this research, came from the necessity of solving a problem that is growing more and more in our society: social anxiety. Social anxiety is reportedly being rising since the Covid-19 isolation, and it is a problem that is affecting personal and professional lives around the world.

Multimodal emotion analysis can contribute to social and therapeutic applications, helping users become aware of their emotional communication patterns, monitor their progress, and adapt behavior in professional or interpersonal contexts. The study's findings are intended to be applied first in the context of preparing oral/public speaking presentations for work or academia, and then be extended to mental health support and virtual communication tools.

5.3 Challenges and Future Directions

This project enables users to reflect on their emotional communication, monitor their progress, and make informed adjustments in future interactions. Despite the work already done, more directions should be followed in order to scale this system.

Achieving stable and low-latency multimodal inference remains computationally demanding, particularly in early fusion, when fusing high-dimensional visual and acoustic data or performing real time emotion recognition. It would be useful udapting the system for smartphones or wearable devices demands lightweight architectures and energy-efficient inference pipelines.

The visual model's performance drops under motion blur, head rotation, or partial occlusion. Robust tracking techniques could help mitigate these issues.

Current models assume a single active speaker. Extending to multi-speaker or group emotion recognition (e.g., meetings, debates) is a natural next step.

It is also important to note that identifying and comparing state-of-the-art methods proved challenging during this research, as existing literature includes many studies focus solely on unimodal emotion recognition, while others explore multimodal setups with varying input combinations and label schemes. This heterogeneity complicates direct comparison and highlights the need for standardized multimodal benchmarks.

Finally, emotions are expressed differently across cultures and contexts, therefore domain adaptation and fairness-aware training could improve accuracy.

Appendix A

Dataset Sources

A.1 Text Emotion Recognition Datasets

Table A.1 Text Emotion Recognition (TER) Datasets.

Dataset	Access Link
ISEAR	https://www.kaggle.com/datasets/faisalsanto007/isear-dataset
GoEmotions	https://www.kaggle.com/datasets/debarshichanda/goemotions
EmoryNLP	https://github.com/emorynlp/emotion-detection
DailyDialog	http://yanran.li/dailydialog
Twitter Emotion Dataset	https://www.kaggle.com/datasets/parulpandey/emotion-dataset
Sentiment140	https://www.kaggle.com/datasets/kazanova/sentiment140

A.2 Speech Emotion Recognition Datasets

Table A.2 Speech Emotion Recognition (SER) Datasets.

Dataset	Access Link
EmoDB	https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb
TESS	https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

A.3 Facial Emotion Recognition Datasets

Table A.3 Facial Emotion Recognition (FER) Datasets.

Dataset	Access Link
CK+	https://www.kaggle.com/datasets/shawon10/ckplus/data https://zenodo.org/records/11221351
FER2013	https://www.kaggle.com/datasets/msambare/fer2013
RAF-DB	http://www.whdeng.cn/RAF/model1.html https://www.kaggle.com/datasets/shuvoalok/raf-db-dataset
AffectNet	https://www.kaggle.com/datasets/mstjebashazida/affectnet
JAFFE	https://zenodo.org/records/14974867
KDEF	https://www.kaggle.com/datasets/chenrich/kdef-database
AFEW	https://users.cecs.anu.edu.au/~few_group/AFEW.html
LFW	https://github.com/amrta-coder/LFW-emotion-dataset

A.4 Multimodal Emotion Recognition Datasets

Table A.4 Multimodal Emotion Recognition (MER) Datasets.

Dataset	Access Link
IEMOCAP	http://sail.usc.edu/iemocap/ https://www.kaggle.com/datasets/dejolilandry/iemocapfullrelease
MELD	https://affective-meld.github.io/ https://www.kaggle.com/datasets/bhandariprakanda/meld-emotion-recognition
SAVEE	http://kahlan.eps.surrey.ac.uk/savee/Download.html Only speech: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
RAVDESS	https://zenodo.org/records/1188976 Audio only: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
CREMA-D	https://www.kaggle.com/datasets/ejlok1/cremad https://github.com/CheyneyComputerScience/CREMA-D/
CMU-MOSEI	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/

Appendix B

GeMAPS Feature Set Description

Table B.1 GeMAPSv01b Functionals Extracted with openSMILE

Idx	Feature Name	Description	Feature Group	Statistic / Unit
1	F0semitoneFrom27.5Hz_sma3nz_amean	Mean pitch in semitone scale (27.5 Hz ref.)	Pitch	Mean / semitone
2	F0semitoneFrom27.5Hz_sma3nz_stddevNorm	Normalized pitch standard deviation	Pitch	Std. / semitone
3	F0semitoneFrom27.5Hz_sma3nz_percentile20.0	20th percentile of fundamental frequency	Pitch	Percentile / semitone
4	F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Median fundamental frequency	Pitch	Percentile / semitone
5	F0semitoneFrom27.5Hz_sma3nz_percentile80.0	80th percentile of fundamental frequency	Pitch	Percentile / semitone
6	F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2	F0 percentile range (0–2)	Pitch	Range / semitone
7	F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope	Mean rising slope of F0	Pitch	Mean / slope

Idx	Feature Name	Description	Feature Group	Statistic / Unit
8	F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	Standard deviation of rising slope	Pitch	Std. / slope
9	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	Mean falling slope of F0	Pitch	Mean / slope
10	F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope	Standard deviation of falling slope	Pitch	Std. / slope
11	loudness_sma3_amean	Mean perceptual loudness	Loudness	Mean / dB
12	loudness_sma3_stddevNorm	Normalized loudness standard deviation	Loudness	Std. / dB
13	loudness_sma3_percentile20.0	20th percentile loudness	Loudness	Percentile / dB
14	loudness_sma3_percentile50.0	Median loudness	Loudness	Percentile / dB
15	loudness_sma3_percentile80.0	80th percentile loudness	Loudness	Percentile / dB
16	loudness_sma3_pctlrange0-2	Loudness percentile range (0–2)	Loudness	Range / dB
17	loudness_sma3_meanRisingSlope	Mean slope during loudness increase	Loudness	Mean / slope
18	loudness_sma3_stddevRisingSlope	Standard deviation of rising slope	Loudness	Std. / slope
19	loudness_sma3_meanFallingSlope	Mean slope during loudness decrease	Loudness	Mean / slope
20	loudness_sma3_stddevFallingSlope	SD of loudness falling slope	Loudness	Std. / slope
21	jitterLocal_sma3nz_amean	Mean local jitter (F0 cycle-to-cycle variation)	Voice Quality	Mean / ratio
22	jitterLocal_sma3nz_stddevNorm	SD of local jitter	Voice Quality	Std. / ratio
23	shimmerLocaldB_sma3nz_amean	Mean local shimmer (amplitude variation)	Voice Quality	Mean / dB

Idx	Feature Name	Description	Feature Group	Statistic / Unit
24	shimmerLocaldB_sma3nz_stddevNorm	SD of shimmer	Voice Quality	Std. / dB
25	HNRdBACF_sma3nz_amean	Mean harmonic-to-noise ratio	Voice Quality	Mean / dB
26	HNRdBACF_sma3nz_stddevNorm	SD of HNR	Voice Quality	Std. / dB
27	logRelF0-H1-H2_sma3nz_amean	Mean log difference between F0, H1, and H2 harmonics	Spectral Balance	Mean / log-ratio
28	logRelF0-H1-H2_sma3nz_stddevNorm	SD of log difference between F0, H1, and H2	Spectral Balance	Std. / log-ratio
29	logRelF0-H1-A3_sma3nz_amean	Mean log difference between F0, H1, and A3	Spectral Balance	Mean / log-ratio
30	logRelF0-H1-A3_sma3nz_stddevNorm	SD of log difference between F0, H1, and A3	Spectral Balance	Std. / log-ratio
31	F1frequency_sma3nz_amean	Mean first formant frequency	Formants	Mean / Hz
32	F1frequency_sma3nz_stddevNorm	SD of F1 frequency	Formants	Std. / Hz
33	F1bandwidth_sma3nz_amean	Mean F1 bandwidth	Formants	Mean / Hz
34	F1bandwidth_sma3nz_stddevNorm	SD of F1 bandwidth	Formants	Std. / Hz
35	F1amplitudeLogRelF0_sma3nz_amean	Mean log amplitude ratio between F1 and F0	Formants	Mean / log-ratio
36	F1amplitudeLogRelF0_sma3nz_stddevNorm	SD of amplitude ratio F1/F0	Formants	Std. / log-ratio
37	F2frequency_sma3nz_amean	Mean second formant frequency	Formants	Mean / Hz
38	F2frequency_sma3nz_stddevNorm	SD of F2 frequency	Formants	Std. / Hz
39	F2amplitudeLogRelF0_sma3nz_amean	Mean amplitude ratio between F2 and F0	Formants	Mean / log-ratio
40	F2amplitudeLogRelF0_sma3nz_stddevNorm	SD of amplitude ratio F2/F0	Formants	Std. / log-ratio

Idx	Feature Name	Description	Feature Group	Statistic / Unit
41	F3frequency_sma3nz_amean	Mean third formant frequency	Formants	Mean / Hz
42	F3frequency_sma3nz_stddevNorm	SD of F3 frequency	Formants	Std. / Hz
43	F3amplitudeLogRelF0_sma3nz_amean	Mean amplitude ratio between F3 and F0	Formants	Mean / log-ratio
44	F3amplitudeLogRelF0_sma3nz_stddevNorm	SD of amplitude ratio F3/F0	Formants	Std. / log-ratio
45	alphaRatioV_sma3nz_amean	Mean spectral alpha ratio (vocalized)	Spectral Tilt	Mean / dB
46	alphaRatioV_sma3nz_stddevNorm	SD of alpha ratio (vocalized)	Spectral Tilt	Std. / dB
47	hammarbergIndexV_sma3nz_amean	Mean Hammarberg index (vocalized)	Spectral Tilt	Mean / dB
48	hammarbergIndexV_sma3nz_stddevNorm	SD of Hammarberg index (vocalized)	Spectral Tilt	Std. / dB
49	slopeV0-500_sma3nz_amean	Spectral slope from 0–500 Hz	Spectral Tilt	Mean / dB/oct
50	slopeV0-500_sma3nz_stddevNorm	SD of spectral slope 0–500 Hz	Spectral Tilt	Std. / dB/oct
51	slopeV500-1500_sma3nz_amean	Spectral slope from 500–1500 Hz	Spectral Tilt	Mean / dB/oct
52	slopeV500-1500_sma3nz_stddevNorm	SD of spectral slope 500–1500 Hz	Spectral Tilt	Std. / dB/oct
53	alphaRatioUV_sma3nz_ amean	Mean alpha ratio (unvoiced)	Spectral Tilt	Mean / dB
54	hammarbergIndexUV_sma3nz_amean	Mean Hammarberg index (unvoiced)	Spectral Tilt	Mean / dB
55	slopeUV0-500_sma3nz_amean	Spectral slope 0–500 Hz (unvoiced)	Spectral Tilt	Mean / dB/oct
56	slopeUV500-1500_sma3nz_amean	Spectral slope 500–1500 Hz (unvoiced)	Spectral Tilt	Mean / dB/oct

Idx	Feature Name	Description	Feature Group	Statistic / Unit
57	loudnessPeaksPerSec	Loudness peaks per second	Temporal	Count / s
58	VoicedSegmentsPerSec	Voiced segments per second	Temporal	Count / s
59	MeanVoicedSegmentLengthSec	Mean voiced segment length	Temporal	Mean / s
60	StddevVoicedSegmentLengthSec	SD of voiced segment length	Temporal	Std. / s
61	MeanUnvoicedSegmentLength	Mean unvoiced segment length	Temporal	Mean / s
62	StddevUnvoicedSegmentLength	SD of unvoiced segment length	Temporal	Std. / s

Bibliography

[1] Robert Plutchik. Plutchik wheel of emotions. https://en.wikipedia.org/wiki/File:Plutchik-wheel.svg, 2011. Public domain, Wikimedia Commons.

[2] Phillip R. Shaver, Joseph Schwartz, Daniel Kirson, and Charlotte O'Connor. Emotion structure based on shaver et al. (1987, p. 1067) [figure], 1987. URL https://www.researchgate.net/publication/330264296/figure/fig1/AS:961493298343940@1606249274436/Emotion-structure-based-on-Shaver-et-al-1987-p-1067.png. Retrieved from ResearchGate figure link.

[3] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980. doi: 10.1037/h0077714.

[4] Vinay Bettadapura. Face expression recognition and analysis: The state of the art, 2012. URL https://arxiv.org/abs/1203.6722.

[5] Ze-Nian Li, Mark Drew, and Jiangchuan Liu. *Fundamentals of Multimedia*. 2004. ISBN 978-3-030-62123-0. doi: 10.1007/978-3-030-62124-7.

[6] Yuki Matsushita, Nori Yasumatsu, Yuki Suzuki, and Yoshiko Matsumoto. A study on hikikomori and its implications for japanese society. *Journal of Humanities and Social Sciences (JHASS)*, 5:121–129, 12 2023. doi: 10.36079/lamintang.jhass-0503.453.

[7] Mustafa Gursesli, Sara Lombardi, Mirko Duradoni, Leonardo Bocchi, Andrea Guazzini, and Antonio lanatà. Facial emotion recognition (fer) through custom lightweight cnn model: Performance evaluation in public datasets. *IEEE Access*, PP:1–1, 01 2024. doi: 10.1109/ACCESS.2024.3380847.

[8] Zhicheng Zhang, Lijuan Wang, and Jufeng Yang. Weakly supervised video emotion detection and prediction via cross-modal temporal erasing network. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18888–18897, 2023. doi: 10.1109/CVPR52729.2023.01811. URL https://openaccess.thecvf.com/

content/CVPR2023/papers/Zhang_Weakly_Supervised_Video_Emotion_Detection_and_Prediction_via_Cross-Modal_Temporal_CVPR_2023_paper.pdf.

[9] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. Analysis of emotional speech—a review. *Toward Robotic Socially Believable Behaving Systems - Volume I : Modeling Emotions*, pages 205–238, 2016. doi: 10.1007/978-3-319-31056-5_11. URL https://doi.org/10.1007/978-3-319-31056-5_11.

[10] Deloitte. Deloitte truevoice, 2024. URL https://www.deloitte.com/uk/en/products/truevoice.html. Accessed: 2025-10-19.

[11] Kairos. Kairos, 2012. URL https://www.kairos.com/.

[12] Huan-Chung Li, Telung Pan, Man-Hua Lee, and Hung-Wen Chiu. Make patient consultation warmer: A clinical application for speech emotion recognition. *Applied Sciences*, 11:4782, 05 2021. doi: 10.3390/app11114782.

[13] Simone Hantke, Hesam Sagha, Nicholas Cummins, and Björn Schuller. Emotional speech of mentally and physically disabled individuals: Introducing the emotass database and first findings. *Interspeech 2017*, pages 3137–3141, 2017. doi: 10.21437/Interspeech.2017-409. URL https://doi.org/10.21437/Interspeech.2017-409.

[14] Nishargo Nigar. Speech emotion recognition using cnn and its use case in digital healthcare. Master's thesis, Hamburg University of Technology, 06 2024.

[15] iMotions. iMotions, 2005. URL https://imotions.com/about-us/.

[16] Smart Eye. Affectiva, 2009. URL https://www.affectiva.com/.

[17] Eyeris. Eyeris, 2013. URL https://www.eyeris.ai/.

[18] Brainchip. Brainchip + NVISO, 2024. URL https://brainchip.com/brainchip-nviso-emotion-detection-demo/.

[19] Microsoft. Cognitive-Emotion-Python, 2021. URL https://github.com/microsoft/Cognitive-Emotion-Python/.

[20] European Parliament and Council of the European Union. Artificial intelligence act (regulation (eu) 2024/1689). Official Journal of the European Union, L 168, 12 July 2024, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[21] Nessa Lynch. Facial recognition technology in policing and security—case studies in regulation. *Laws*, 13(3), 2024. ISSN 2075-471X. doi: 10.3390/laws13030035. URL https://www.mdpi.com/2075-471X/13/3/35.

[22] Charles Darwin. *The Expression of the Emotions in Man and Animals*. John Murray, London, 1872.

[23] William James. What is an emotion? *Mind*, 9(34):188–205, 1884. ISSN 00264423, 14602113. URL http://www.jstor.org/stable/2246769.

[24] Paul Ekman, Wallace V. Friesen, and Phoebe C. Ellsworth. *Emotion in the Human Face: Guidelines for Research and an Integration of Findings*. Pergamon General Psychology Series. Pergamon Press, New York, 1972. ISBN 9780080166438.

[25] Albert Mehrabian and James A. Russell. *An Approach to Environmental Psychology*. The MIT Press, Cambridge, MA, 1974. ISBN 0262130904.

[26] David Rubin and Jennifer Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory (Hove, England)*, 17:802–8, 09 2009. doi: 10.1080/09658210903130764.

[27] H. Schlosberg. The description of facial expressions in terms of two pages. *Journal of Experimental Psychology*, pages 229–237, 1952. doi: https://doi.org/10.1037/h0055778.

[28] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219–235, 1985.

[29] Dimitrios Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48:1162–1181, 09 2006. doi: 10.1016/j.specom.2006.04.003.

[30] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21, 03 2018. doi: 10.1007/s10772-018-9491-z.

[31] Ruhul Amin Khalil, Edward Jones, Mohammad Babar, Tariqullah Jan, Mohammad Zafar, and Thamer Alhussain. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, PP:1–1, 08 2019. doi: 10.1109/ACCESS.2019.2936124.

[32] Kaggle. Kagglehub repository, 2024. URL https://github.com/Kaggle/kagglehub. Accessed: October 20, 2025.

[33] Wallbott H. G. Scherer, K. R. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, pages 310–328, 1994. URL https://doi.org/10.1037/0022-3514.66.2.310.

[34] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547, 2020. URL https://arxiv.org/abs/2005.00547.

[35] Google Research. Goemotions dataset repository, 2021. URL https://github.com/google-research/google-research/tree/master/goemotions. Accessed: 2025-10-19.

[36] Sayyed M. Zahiri and Jinho D. Choi. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299, 2017. URL http://arxiv.org/abs/1708.04299.

[37] Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Li. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics, 2025. URL https://arxiv.org/abs/2403.07260.

[38] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *CoRR*, abs/1710.03957, 2017. URL http://arxiv.org/abs/1710.03957.

[39] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://www.aclweb.org/anthology/D18-1404.

[40] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150, 01 2009.

[41] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. A database of german emotional speech. volume 5, pages 1517–1520, 09 2005. doi: 10.21437/Interspeech.2005-446.

[42] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

[43] University of Toronto. Toronto emotional speech set (tess), 2010. URL https://utoronto.scholaris.ca/collections/036db644-9790-4ed0-90cc-be1dfb8a4b66. Accessed: 2025-10-19.

[44] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010. doi: 10.1109/CVPRW.2010.5543262.

[45] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. doi: 10.1109/CVPR.2017.277. RAF-DB: Real-world Affective Faces Database.

[46] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985, 2017. URL http://arxiv.org/abs/1708.03985.

[47] Michael J. Lyons. "excavating ai" re-excavated: Debunking a fallacious account of the JAFFE dataset. *CoRR*, abs/2107.13998, 2021. URL https://arxiv.org/abs/2107.13998.

[48] Michael J. Lyons, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets (IVC special issue). *CoRR*, abs/2009.05938, 2020. URL https://arxiv.org/abs/2009.05938.

[49] Flykt A. Öhman A. Lundqvist, D. The karolinska directed emotional faces - kdef. *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.*, 1998. doi: 10.1037/t27732-000.

[50] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008. doi: 10.1007/s10579-008-9076-6.

[51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *CoRR*, abs/1810.02508, 2018. URL http://arxiv.org/abs/1810.02508.

[52] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao K. Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. *CoRR*, abs/1802.08379, 2018. URL http://arxiv.org/abs/1802.08379.

[53] Philip Jackson and Sana ul haq. Surrey audio-visual expressed emotion (savee) database, 04 2011.

[54] Steven R. Livingstone and Frank 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*, 13(5):e0196391, 2018. doi: 10.1371/journal.pone.0196391. URL https://zenodo.org/record/1188976.

[55] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

[56] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1208.

[57] Dias Issa, M. Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 05 2020. doi: 10.1016/j.bspc.2020.101894.

[58] Smith K. Khare, Victoria Blanes-Vidal, Esmaeil S. Nadimi, and U. Rajendra Acharya. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, 102:102019, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102019. URL https://www.sciencedirect.com/science/article/pii/S1566253523003354.

[59] Seyed Esfahani and Mehdi Adda. Classical machine learning and large models for text-based emotion recognition. *Procedia Computer Science*, 241:77–84, 01 2024. doi: 10.1016/j.procs.2024.08.013.

[60] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[61] M.F. Porter. An algoritm for suffix stripping. *Program*, 14:130–137, 07 2006. doi: 10.1108/00330330610681286.

[62] NLTK Project. Natural language toolkit (nltk) api. URL https://www.nltk.org/api/nltk.html. Accessed: 2025-10-19.

[63] Explosion AI. spacy api documentation. URL https://spacy.io/api. Accessed: 2025-10-19.

[64] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

[65] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

[66] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654, 2020. URL https://arxiv.org/abs/2006.03654.

[67] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018. URL http://arxiv.org/abs/1808.06226.

[68] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, abs/1804.10959, 2018. URL http://arxiv.org/abs/1804.10959.

[69] Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. XLM-T: A multilingual language model toolkit for twitter. *CoRR*, abs/2104.12250, 2021. URL https://arxiv.org/abs/2104.12250.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

[71] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, March 2003. ISSN 1532-4435.

[72] Daniel Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 11 1999. doi: 10.1038/44565.

[73] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. pages 1–12, 01 2013.

[74] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. volume 14, pages 1532–1543, 01 2014. doi: 10.3115/v1/D14-1162.

[75] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

[76] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

[77] James Pennebaker, Martha Francis, and Roger Booth. Linguistic inquiry and word count (liwc). 01 1999.

[78] Mohammad, Saif. Nrc emotion lexicon (emolex). URL https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. Accessed: 2025-10-19.

[79] Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *CoRR*, abs/2106.01071, 2021. URL https://arxiv.org/abs/2106.01071.

[80] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *CoRR*, abs/2108.12009, 2021. URL https://arxiv.org/abs/2108.12009.

[81] Lorenzo Vaiani, Luca Cagliero, and Paolo Garza. Emotion recognition from videos using multimodal large language models. *Future Internet*, 16(7), 2024. ISSN 1999-5903. doi: 10.3390/fi16070247. URL https://www.mdpi.com/1999-5903/16/7/247.

[82] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models, 2024. URL https://arxiv.org/abs/2309.11911.

[83] Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Li. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics, 2025. URL https://arxiv.org/abs/2403.07260.

[84] Rafael Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *T. Affective Computing*, 1:18–37, 01 2010. doi: 10.1109/T-AFFC.2010.1.

[85] Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. Deep emotion recognition in textual conversations: A survey, 2024. URL https://arxiv.org/abs/2211.09172.

[86] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017. URL http://arxiv.org/abs/1705.09406.

[87] Berkehan Akçay and Kaya Oguz. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 01 2020. doi: 10.1016/j.specom.2019.12.001.

[88] Kuan-Yu Chen and Kun-Mao Chao. Optimal algorithms for locating the longest and shortest segments satisfying a sum or an average constraint. *Information Processing Letters*, 96(6):197–201, 2005. ISSN 0020-0190. doi: https://doi.org/10.1016/j.ipl.2005.08.006. URL https://www.sciencedirect.com/science/article/pii/S0020019005002206.

[89] Florian Eyben, Martin Wöllmer, and Björn Schuller. opensmile – the munich versatile and fast open-source audio feature extractor. pages 1459–1462, 01 2010. doi: 10.1145/1873951.1874246.

[90] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. pages 18–24, 01 2015. doi: 10.25080/Majora-7b98e3ed-003.

[91] Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot Int*, 5: 341–347, 01 2001.

[92] Florian Eyben, Martin Wollmer, and Björn Schuller. Openear - introducing the munich open-source emotion and affect recognition toolkit. pages 1 – 6, 10 2009. doi: 10.1109/ACII.2009.5349350.

[93] Milana Bojanic, Vlado Delić, and Alexey Karpov. Call redistribution for a call center based on speech emotion recognition. *Applied Sciences*, 10:4653, 07 2020. doi: 10.3390/app10134653.

[94] Florian Eyben, Klaus R. Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:190–202, 2016. URL https://api.semanticscholar.org/CorpusID:14486649.

[95] Anjali Bhavan, Pankaj Chauhan, Hitkul, and Rajiv Ratn Shah. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184:104886, 08 2019. doi: 10.1016/j.knosys.2019.104886.

[96] Moataz Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 03 2011. doi: 10.1016/j.patcog.2010.09.020.

[97] Chandni, Garima Vyas, Malay Kishore Dutta, Kamil Riha, and Jiri Prinosil. An automatic emotion recognizer using mfccs and hidden markov models. In *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 320–324, 2015. doi: 10.1109/ICUMT.2015.7382450.

[98] Tin Nwe, S.W. Foo, and Liyanage De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 11 2003. doi: 10.1016/S0167-6393(03)00099-2.

[99] Iker Luengo, Eva Navas, Inmaculada Hernáez, and Jon Sánchez. Automatic emotion recognition using prosodic parameters. pages 493–496, 09 2005. doi: 10.21437/Interspeech.2005-324.

[100] Yashpalsing Chavhan, Manikrao Dhore, and Yesaware Pallavi. Speech emotion recognition using support vector machines. *International Journal of Computer Applications*, 1, 02 2010. doi: 10.1007/978-3-642-21402-8_35.

[101] Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, and Rajesh Kumar Muthu. Speech emotion recognition using support vector machine, 2020. URL https://arxiv.org/abs/2002.07590.

[102] Jarosław Cichosz and Krzysztof Slot. Emotion recognition in speech signal using emotion-extracting binary decision trees. 01 2007.

[103] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 15:1–32, 12 2019. doi: 10.1145/3363560.

[104] Arya Aftab, Alireza Morsali, Shahrokh Ghaemmaghami, and Benoit Champagne. Lightsernet: A lightweight fully convolutional neural network for speech emotion recognition. 10 2021. doi: 10.48550/arXiv.2110.03435.

[105] Mustaqeem and Soonil Kwon. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, 167:114177, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.114177. URL https://www.sciencedirect.com/science/article/pii/S0957417420309131.

[106] Ilkhomjon Pulatov, Rashid Oteniyazov, Fazliddin Makhmudov, and Young-Im Cho. Enhancing speech emotion recognition using dual feature extraction encoders. *Sensors*, 23(14), 2023. ISSN 1424-8220. doi: 10.3390/s23146640. URL https://www.mdpi.com/1424-8220/23/14/6640.

[107] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019. doi: 10.1109/ACCESS.2019.2938007.

[108] Yong Wang, Cheng Lu, Hailun Lian, Yan Zhao, Björn Schuller, Yuan Zong, and Wenming Zheng. Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition, 2024. URL https://arxiv.org/abs/2401.10536.

[109] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. Speechformer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:775–788, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2023.3235194. URL http://dx.doi.org/10.1109/TASLP.2023.3235194.

[110] Chunjun Zheng, Chunli Wang, and Jia Ning. An ensemble model for multi-level speech emotion recognition. *Applied Sciences*, 10:205, 12 2019. doi: 10.3390/app10010205.

[111] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 01 2019. doi: 10.1016/j.bspc.2018.08.035.

[112] C Kumar, Advaith Maharana, Srinath Krishnan, Sannidhi Hanuma, Jyothish Lal G, and Vinayakumar Ravi. *Speech Emotion Recognition Using CNN-LSTM and Vision Transformer*, pages 86–97. 03 2023. ISBN 978-3-031-27498-5. doi: 10.1007/978-3-031-27499-2_8.

[113] Zhang Kexin and Liu Yunxiang. Speech emotion recognition based on transfer emotion-discriminative features subspace learning. *IEEE Access*, PP:1–1, 01 2023. doi: 10.1109/ACCESS.2023.3282982.

[114] Sitong Zhou and Homayoon Beigi. A transfer learning method for speech emotion recognition from automatic speech recognition, 2020. URL https://arxiv.org/abs/2008.02863.

[115] Daria Diatlova, Anton Udalov, Vitalii Shutov, and Egor Spirin. Adapting wavlm for speech emotion recognition, 2024. URL https://arxiv.org/abs/2405.04485.

[116] Fabian Ritter-Gutierrez, Kuan-Po Huang, Jeremy H. M Wong, Dianwen Ng, Hung yi Lee, Nancy F. Chen, and Eng Siong Chng. Dataset-distillation generative model for speech emotion recognition, 2024. URL https://arxiv.org/abs/2406.02963.

[117] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Alexandros Potamianos, and Shrikanth Narayanan. Data augmentation using gans for speech emotion recognition. pages 171–175, 09 2019. doi: 10.21437/Interspeech.2019-2561.

[118] Amir Shirian and Tanaya Guha. Compact graph architecture for speech emotion recognition. *CoRR*, abs/2008.02063, 2020. URL https://arxiv.org/abs/2008.02063.

[119] Siyuan Shen, Yu Gao, Feng Liu, Hanyang Wang, and Aimin Zhou. Emotion neural transducer for fine-grained speech emotion recognition. pages 10111–10115, 04 2024. doi: 10.1109/ICASSP48485.2024.10446974.

[120] Shuaiqi Chen, Xiaofen Xing, Weibin Zhang, Weidong Chen, and Xiangmin Xu. Dwformer: Dynamic window transformer for speech emotion recognition, 2023. URL https://arxiv.org/abs/2303.01694.

[121] Tiantian Feng, Rajat Hebbar, and Shrikanth Narayanan. Trustser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition, 2023. URL https://arxiv.org/abs/2305.11229.

[122] Andre Lopes, Edilson Aguiar, Alberto De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61, 07 2016. doi: 10.1016/j.patcog.2016.07.026.

[123] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *CoRR*, abs/1604.03225, 2016. URL http://arxiv.org/abs/1604.03225.

[124] Felipe Canal, Tobias Müller, Jhennifer Matias, Gustavo Scotton, Antonio Junior, Eliane Pozzebon, and Antonio Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582, 10 2021. doi: 10.1016/j.ins.2021.10.005.

[125] Deepak Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120, 04 2019. doi: 10.1016/j.patrec.2019.01.008.

[126] Maja Pantic and Léon Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1424 – 1445, 01 2001. doi: 10.1109/34.895976.

[127] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. volume 1, pages I–511, 02 2001. ISBN 0-7695-1272-0. doi: 10.1109/CVPR.2001.990517.

[128] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

[129] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. URL http://arxiv.org/abs/1604.02878.

[130] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou.

Retinaface: Single-stage dense face localisation in the wild. *CoRR*, abs/1905.00641, 2019. URL http://arxiv.org/abs/1905.00641.

[131] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, January 1995. ISSN 1077-3142. doi: 10.1006/cviu.1995.1004. URL https://doi.org/10.1006/cviu.1995.1004.

[132] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. *CoRR*, abs/1706.01789, 2017. URL http://arxiv.org/abs/1706.01789.

[133] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *CoRR*, abs/1902.09212, 2019. URL http://arxiv.org/abs/1902.09212.

[134] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *SIGGRAPH'99 Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 09 2002. doi: 10.1145/311535.311556.

[135] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. doi: 10.1109/TPAMI.2002.1017623.

[136] Khadija Slimani, Mohamed Kas, Youssef El Merabet, Rochdi Messoussi, and Yassine Ruichek. Facial emotion recognition: A comparative analysis using 22 lbp variants. pages 88–94, 03 2018. doi: 10.1145/3177148.3180092.

[137] Taskeed Jabid, Md Kabir, and Oksam Chae. Robust facial expression recognition based on local directional pattern. *ETRI Journal*, 32, 10 2010. doi: 10.4218/etrij.10.1510.0132.

[138] Taskeed Jabid, Md. Hasanul Kabir, and Oksam Chae. Local directional pattern (ldp) for face recognition. In *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pages 329–330, 2010. doi: 10.1109/ICCE.2010.5418801.

[139] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998. doi: 10.1109/AFGR.1998.670949.

[140] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005. doi: 10.1109/CVPR.2005.177.

[141] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018. doi: 10.1109/FG.2018.00019.

[142] Tadas Baltrušaitis. Openface: Facial behavior analysis toolkit. URL `https://github.com/TadasBaltrusaitis/OpenFace`. Accessed: 2025-10-19.

[143] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and S. Baker. Multi-pie. 12 2013. doi: 10.1109/AFGR.2008.4813399.

[144] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

[145] Research Publication. Very deep convolutional networks for large-scale image recognition vol 12 issue 08. *SSRN Electronic Journal*, 12:301–307, 08 2012.

[146] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

[147] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[148] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.

[149] Ali Mollahosseini, Behzad Hasani, and Mohammad Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP, 08 2017. doi: 10.1109/TAFFC.2017.2740923.

[150] Wafa Mellouk and Handouzi Wahida. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694, 01 2020. doi: 10.1016/j.procs.2020.07.101.

[151] Ira Cohen, Nicu Sebe, Ashutosh Garg, and Lawrence Chen. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 07 2003. doi: 10.1016/S1077-3142(03)00081-X.

[152] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. pages 435–442, 11 2015. doi: 10.1145/2818346.2830595.

[153] Yang Bo, Jianming Wu, and Gen Hattori. Facial expression recognition with the advent of face masks, 11 2020.

[154] Diego Mushfieldt, Mehrdad Ghaziasgar, and James Connan. Robust facial expression recognition in the presence of rotation and partial occlusion. pages 186–193, 10 2013. doi: 10.1145/2513456.2513493.

[155] Mehmet Emin Konuk and Erdal Kılıç. Efficientfer: Efficientnetv2 based deep learning approach for facial expression recognition. In *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, pages 1–7, 2025. doi: 10.1109/ICHORA65333.2025.11017006.

[156] Ce Zheng, Matias Mendieta, and Chen Chen. Poster: A pyramid cross-fusion transformer network for facial expression recognition, 2023. URL https://arxiv.org/abs/2204.04083.

[157] Rui Sun, Zhaoli Zhang, and Hai Liu. Fcca: Fast center consistency attention for facial expression recognition. *Electronics*, 14(6), 2025. ISSN 2079-9292. doi: 10.3390/electronics14061057. URL https://www.mdpi.com/2079-9292/14/6/1057.

[158] Vladimir Chernykh, Grigoriy Sterling, and Pavel Prihodko. Emotion recognition from speech with recurrent neural networks. *CoRR*, abs/1701.08071, 2017. URL http://arxiv.org/abs/1701.08071.

[159] Lei Pang and Chong-Wah Ngo. Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos. pages 619–622, 06 2015. doi: 10.1145/2671188.2749400.

[160] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *2008 Tenth IEEE International Symposium on Multimedia*, pages 250–257, 2008. doi: 10.1109/ISM.2008.40.

[161] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. pages 2362–2365, 09 2010. doi: 10.21437/Interspeech.2010-646.

[162] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. pages 3687–3691, 10 2013. doi: 10.1109/ICASSP.2013.6638346.

[163] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:1359–1367, 04 2020. doi: 10.1609/aaai.v34i02.5492.

[164] José Salas-Cáceres, Javier Lorenzo-Navarro, David Freire-Obregón, and Modesto Castrillón Santana. Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. *Multimedia Tools and Applications*, 84:27327–27343, 09 2024. doi: 10.1007/s11042-024-20227-6.

[165] Prabhav Singh, Ridam Srivastava, K. Rana, and Vineet Kumar. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, 229:107316, 10 2021. doi: 10.1016/j.knosys.2021.107316.

[166] Lili Guo, Longbiao Wang, Jianwu Dang, Yahui Fu, Jiaxing Liu, and Shifei Ding. Emotion recognition with multimodal transformer fusion framework based on acoustic and lexical information. *IEEE MultiMedia*, 29:1–1, 04 2022. doi: 10.1109/MMUL.2022.3161411.

[167] Haytham Fayek, Margaret Lech, and L. Cavedon. Towards real-time speech emotion recognition using deep neural networks. 12 2015. doi: 10.1109/ICSPCS.2015.7391796.

[168] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances, 2024. URL https://arxiv.org/abs/2407.21315.

[169] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning, 2024. URL https://arxiv.org/abs/2406.11161.

[170] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations, 2024. URL https://arxiv.org/abs/2310.11374.

[171] Shanmin Wang, Chengguang Liu, and Qingshan Liu. Multi-modality collaborative learning for sentiment analysis, 2025. URL https://arxiv.org/abs/2501.12424.

[172] Peihao Xiang, Kaida Wu, and Ou Bai. Mtcae-dfer: Multi-task cascaded autoencoder for dynamic facial expression recognition, 2025. URL https://arxiv.org/abs/2412.18988.

[173] Peihao Xiang, Chaohao Lin, Kaida Wu, and Ou Bai. Multimae-der: Multimodal masked autoencoder for dynamic emotion recognition. In *2024 14th International Conference on Pattern Recognition Systems (ICPRS)*, page 1–7. IEEE, July 2024. doi: 10.1109/icprs62101.2024.10677820. URL http://dx.doi.org/10.1109/ICPRS62101.2024.10677820.

[174] Xuecheng Wu, Heli Sun, Junxiao Xue, Ruofan Zhai, Xiangyan Kong, Jiayu Nie, and Liang He. emotions: A large-scale dataset for emotion recognition in short videos. *arXiv preprint arXiv:2311.17335*, 2023.

[175] Qinglan Wei, Xuling Huang, and Yuan Zhang. Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference, 2022. URL https://arxiv.org/abs/2209.10170.

[176] Didan Deng, Zhaokang Chen, Yuqian Zhou, and Bert Shi. Mimamo net: Integrating micro- and macro-motion for video emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2621–2628, 04 2020. doi: 10.1609/aaai.v34i03.5646.

[177] Whisper API. Which *Whisper* model should i choose? https://whisper-api.com/blog/models/, March 2025. Accessed: 2025-10-21.