Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: elektronica-ICT

Masterthesis

Implementing Emotion Detection from Speech in a Commercial Robot for Psychological Assessment of Elderly People: A Comparative Study of Python-based Approaches and Existing Solutions

Ismael Warnants

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: elektronica-ICT

PROMOTOR :
dr. Nikolaos TSIOGKAS

PROMOTOR :
Prof. dr. Joaquin Francisco ROCA GONZÁLEZ

COPROMOTOR :
Prof. Francisco José ORTIZ ZARAGOZA

Gezamenlijke opleiding UHasselt en KU Leuven

UHASSELT KU LEUVEN

Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt

UHASSELT KU LEUVEN

2022
2023

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: elektronica-ICT

Masterthesis

Implementing Emotion Detection from Speech in a Commercial Robot for Psychological Assessment of Elderly People: A Comparative Study of Python-based Approaches and Existing Solutions

Ismael Warnants

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: elektronica-ICT

PROMOTOR :

dr. Nikolaos TSIOGKAS

PROMOTOR :

Prof. dr. Joaquin Francisco ROCA GONZÁLEZ

COPROMOTOR :

Prof. Francisco José ORTIZ ZARAGOZA

UHASSELT KU LEUVEN

Table of contents

Preface

This master's thesis was one of the most interesting projects I was part of in my life because I was able to work on a project in Spain with an incredible team working on the ADDIM system. I want to thank all of the team members, especially the team supervisor Prof. Dr. Francisco José Ortiz Zaragoza who dedicated a lot of time to ensure that everything within the project went without problems. Additionally I want to thank my promotors from Spain Prof. Dr. Joaquín Francisco Roca González and from Belgium dr. Nikolaos Tsiogkas for both supporting me a lot with my project. I also want to thank Juan Pedro Serna, a physician, and CEO of E-Doctor, for advising me on this project, which resulted in a better end product.

I also want to thank the international office of KU Leuven, UHasselt, and UPCT for making this Erasmus possible.

Aside from this I also want to thank all of my professors, teachers, family, and friends for their support during the years of my education. It is because of them that I was able to work on this project.

List of tables

List of figures

Abstract in English

In the last ten years, the number of people over 65 has increased 30% in Spain. This trend is anticipated to grow and require more healthcare personnel. To prevent this, people should live longer independently instead of in care homes. The ADDIM system will assist them in living independently. This master's thesis aims to recognize the emotion of the user. This will be part of the mood detection of the user in the ADDIM (Asistencia Domiciliaria Digital Integral para Mayores) system. This is a Digital platform for monitoring older people's health, safety, companionship, and emotional support at home based on robotics, artificial intelligence, and ambient assisted living.

To detect user emotions, the right speech corpus, feature extraction methods, preprocessing methods, and machine learning models have to be selected. Based on the detected emotion, the robot will interact with the user to perform predefined actions. The final mood of the user will be estimated using this output in conjunction with visual feedback and the sensors in the user's home with the ADDIM system.

Three speech corpora are selected with retraining to achieve personalized detection based on the user's previous recordings. In addition, this will ensure that the detection is improved over time, which has yet to be implemented in other research. Finally, the implementation uses dimensional emotion detection instead of discrete emotion detection. This augments the number of detectable emotions.

Abstract in Dutch

In de afgelopen tien jaar is het aantal 65-plussers in Spanje met 30% gestegen. Om dit tegen te gaan zal het ADDIM systeem de gebruiker ondersteunen om zelfstandig te wonen. Deze masterthesis is gericht op het herkennen van de emotie van de gebruiker. Dit zal deel uitmaken van de stemmingsdetectie van de gebruiker in het ADDIM (Asistencia Domiciliaria Digital Integral para Mayores) systeem. Dit is een digitaal platform voor het in de gaten houden van de gezondheid, veiligheid, gezelschap en emotionele ondersteuning van ouderen thuis, gebaseerd op robotica, kunstmatige intelligentie en ambient assisted living.

Om gebruikersemoties te kunnen detecteren, moeten het juiste spraakcorpus, kenmerkextractiemethoden, voorbewerkingsmethoden en machine-learningmodellen worden geselecteerd. Op basis van de gedetecteerde emotie zal de robot de interactie aangaan met de gebruiker om voorgedefinieerde acties uit te voeren. De uiteindelijke stemming van de gebruiker wordt ingeschat door deze output in combinatie met visuele feedback en de sensoren in het huis van de gebruiker met het ADDIM-systeem.

Drie spraak corpora zijn gekozen met hertraining om gepersonaliseerde detectie te bereiken op basis van eerdere opnames van de gebruiker. Bovendien zorgt dit ervoor dat de detectie in de loop van de tijd wordt verbeterd, wat in eerder onderzoek nog niet is geïmplementeerd. Ten slotte gebruikt de implementatie dimensionale emotiedetectie in plaats van discrete emotiedetectie. Dit vergroot het aantal detecteerbare emoties.

1 Introduction

In the last ten years, the number of people over 65 has increased by 30% In Spain. One of the reasons is that “In Spain, life expectancy at birth has improved by 4,14 years from 79,1 years in 2000 to 83,2 years in 2019.” [1]. These statements have multiple implications. On the one hand, the increase in life expectancy is perceived as a positive result of the increased quality of life. On the other hand, this means that more older people will live alone because of slow birth rates, as shown in Figure 1. Also, women have a longer life expectancy than men in Spain. And that is even double for the age class above 85 years old. This trend is also noticeable in the rest of Europe, as shown in Figure 2.

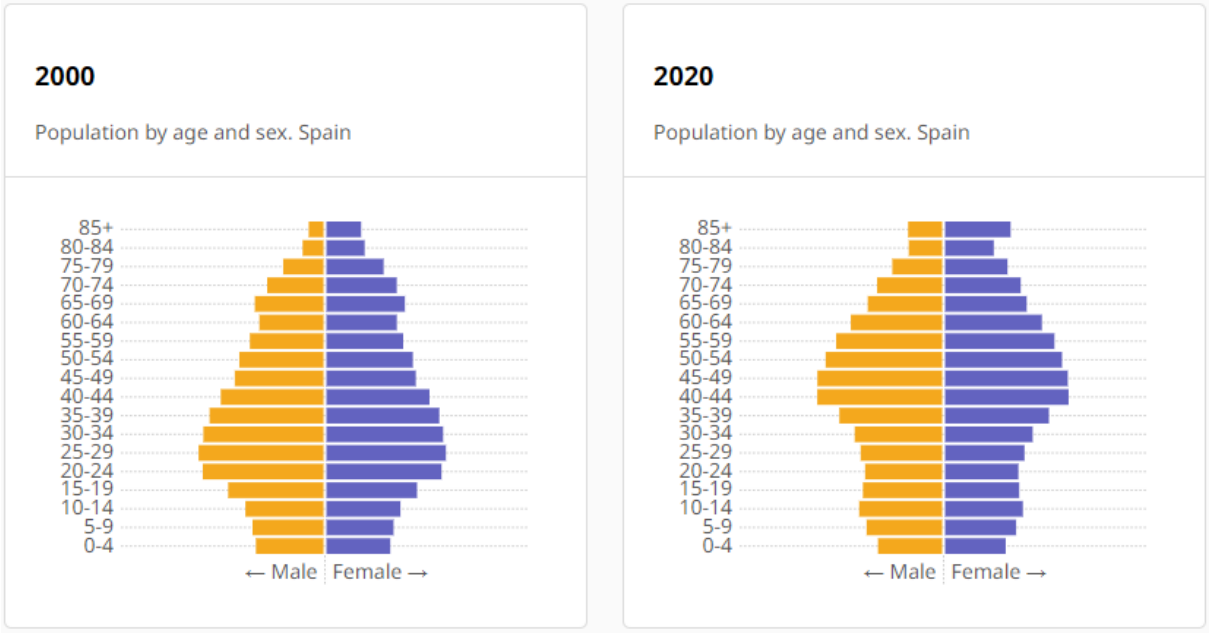

Figure 1 The population by age and sex of Spain in 2000 (left) and 2020 (right) [1]

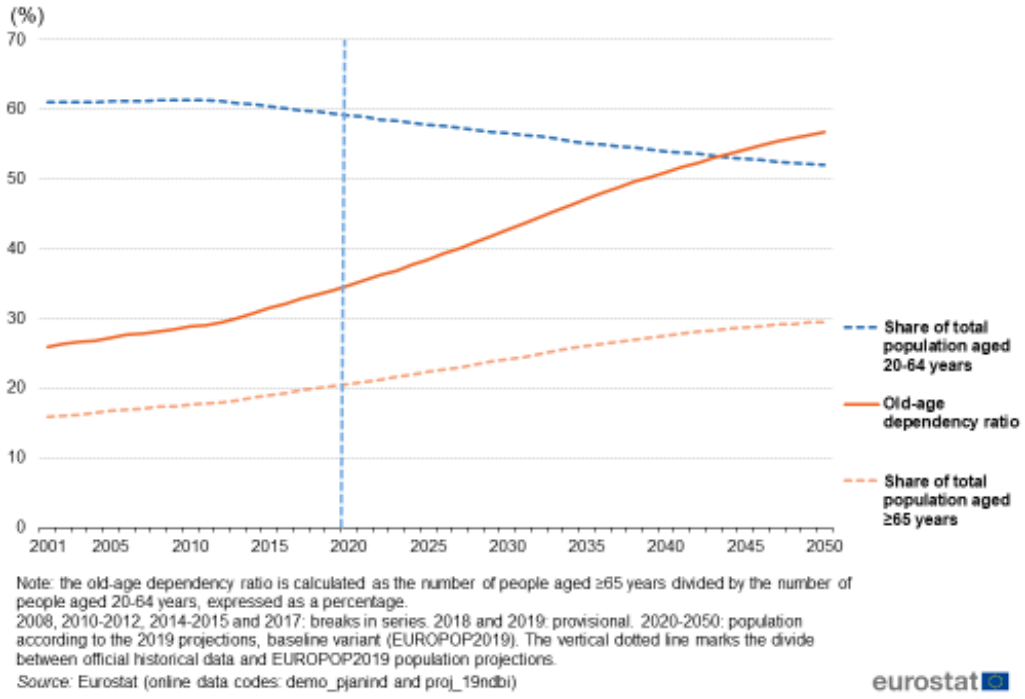

Figure 2 Population structure indicators, EU-27, 2001-2050 [2]

As shown in Figure 3, the growth of women and men above 65 will significantly increase by 2050. This means there will be fewer medical personnel to care for older people.

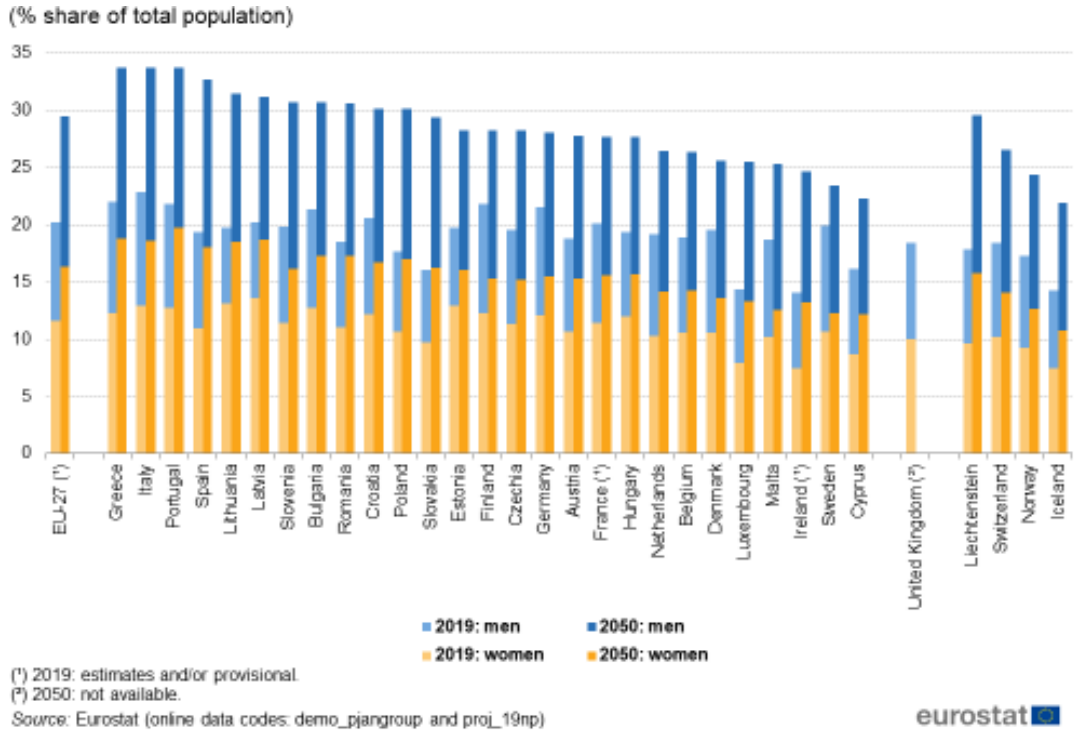

Figure 3 People aged ≥65 years, by sex, 2019 and 2050 [2]

Because of this, the quality of life and the mental health of older people will decrease, which could lead to mental disorders or even suicide. According to the data in Figure 4, Spanish people aged 65 and above accounted for 53% of all reported suicide cases.

Last updated: 2021-07-06

Indicator	Crude suicide rates (per 100 000 population)		
Period	2019		
Location	Both sexes	Male	Female
Spain			
85+ years	13.57	30.53	4.74
75-84 years	16.22	28.01	7.37
65-74 years	10.26	14.94	6.09
55-64 years	10.3	14.92	5.87
45-54 years	10.24	15.07	5.29
35-44 years of age	7.18	10.31	3.99
25-34 years of age	4.87	7.44	2.26
15-24 years	2.71	3.7	1.67

Figure 4 Crude suicide rates (per 100 000 population) in Spain [3]

To support the growing elderly population and prevent this trend, solutions have to be developed to support them to live alone by providing them with personal care. But as mentioned before, there needs to be more medical personnel to take care of this. To overcome this, assistive technologies have been developed and have become more accessible over the years as technology prices have

decreased. Assistive technologies could support older people in everyday tasks and contact human caregivers or emergency services when necessary.

These assistive systems can personalize their care for the elderly user by knowing their emotion. This paper proposes a Speech Emotion Recognition (SER) model to detect emotions from speech using a machine learning model. The speech was chosen because it is a non-invasive method to recognize emotions everywhere in the room. For years, similar principles have been shown to perform Speech Recognition on intelligent home speakers from brands like Google, Apple, and Amazon. Aside from this, it is also more versatile than Facial Emotion Recognition (FER), which requires a camera pointing directly at the face/body. However, privacy concerns must be considered to avoid user data being leaked to the public. This can be achieved by processing everything locally.

This SER model will be part of the ADDIM (Asistencia Domiciliaria Digital Integral para Mayores) system, which is a digital platform for monitoring the health, safety, companionship, and emotional support of older people at home based on robotics, artificial intelligence, and ambient assisted living. This continues the HIMTAE (Heterogeneous Intelligent Multirobot Team for the Assistance of Elderly People) project. Part of the goal of ADDIM is to detect the mood of the elderly user by taking input from the SER model and its sensors. Both projects are a collaboration between e-Doctor, INTEC Sistemas, and a research team at the UPCT (Universidad Polytécnica de Cartagena). E-Doctor is a telemedicine platform with a group of psychologists and doctors behind. INTEC Sistemas is the company that will work with this team to develop the final robot. And the research team at UPCT is developing the core system and leading this project. After development, a pilot project will be tested in Poncemar Foundation in Lorca (Spain).

In the present work, many approaches are proposed to recognize discrete emotions based on many existing databases. This paper will attempt to choose or create the best model to implement into the ADDIM system while considering its requirements. The detection should be finished within five minutes to detect the user's emotional state, simple enough to run on a CPU to save cost, built for long-term use, and with privacy in mind.

The implementation goal is to be a proof-of-concept to check the feasibility of triggering actions in a robot based on the result provided by the SER model. The possible ways to improve the SER model and its real-world limitations will be explored for future research. Aside from this, the aim is to detect emotions in audio fragments instead of continuous speech. There are also some assumptions made. For example, the noise level is considered constant, and the focus is to detect the speech of a single person. Additionally, suppose the Concordance Correlation Coefficient (CCC) is around or above 0.5 [4]. In that case, this is enough to combine this output with the data to train the mood detection that will later be implemented into the ADDIM system.

2 Methodology

The methodology will describe the aspects to consider to create a SER model.

Section 2.1 describes the research and design approaches taken in this paper.

Section 2.2 describes how emotions can be interpreted and the aspects that need to be considered when attempting to interpret emotions.

Section 2.3 describes the steps and considerations when choosing or creating SER models.

Section 2.4 describes the state-of-the-art results of SER that have been achieved.

Section 2.5 describes considerations that need to be taken to apply SER in real-world applications.

Section 2.6 describes related work of emotion recognition in robots and elderly care.

2.1 Research and design approach

The approach in this study consists of two parts, the comparative part and the explorative part. The comparative part looks at the state-of-the-art to determine the approaches that can be taken in the explorative part of this study. The explorative part will take these approaches, combine them and implement them into the study. This is why the explorative part is necessary in this study.

This chapter consists of the comparative part, and the following chapters will explore the fittest solution to be implemented and explain the approach taken to achieve these results.

2.2 Interpreting emotions

Emotion "can be thought of as the emotional 'climate' of the brain" [5], this 'climate' can change often [5]. Interpreting emotions is a mundane task for most people. However, estimating emotions with a robot requires a method to describe the emotion it detects. There are two methods to classify the emotions that can be detected [6].

2.2.1 Emotion classification methods

The first method to classify emotions is to define a discrete or limited set of emotions that will be labeled. These labels are often used to label speech emotion datasets or speech corpora. The available emotions in a speech corpora vary but are often limited to four to eight [6]–[9]. The robot detects emotions consisting of a single emotion or a set of possibilities for all the labeled emotions.

The second method to classify emotions is to define them on multiple dimensions to estimate the closest emotion [4]. This is called the Speech Affective Space Model (SASM) [10]. There are three different dimensions mentioned in the state-of-the-art, which are Valence (Pleasure), Arousal (Activity), and Dominance (Control) [11]:

- Valence describes the amount of "pleasure" expressed in the emotion [11]. This can range from negative (extremely unpleasant) to positive (highly pleasant) [11].
- Arousal or Activity describes the amount of intensity that is expressed in the emotion [11]. This can range from negative (calm, drowsy, or peaceful) to positive (energized, excited, and alert) [11].

- Dominance describes on a particular Valence-Arousal point the amount of control over a situation is expressed in the emotion [11]. This can range from negative (no control) to positive (complete control) [11].

Depending on the detection goal, the number of axes and labeled emotions on them can vary. In state-of-the-art research, the discrete emotions are preferred [7]–[9], however in the ADDIM system, using the Valence and Arousal axes is preferred. In the state-of-the-art of dimensional emotion detection, the amount of emotions labeled on two dimensions varies [4], [6]. Table 1 and Table 2 show the number of emotions that can be mapped. Table 1 is an example of a Recalibrated SASM (rSASM), and Table 2 is an example of a SASM. With an increased amount of possible emotions, the detection accuracy falls [4], [6].

Table 1 Intended output for different emotions [10]

Emotion	Valence Value	Arousal value	Quadrant
Anger	-1	+1	2
Happiness	+1	+1	1
Sadness	-1	-1	3
Neutral	0	0	Center
Calm	+1	-1	4

Table 2 Initialised labels for the circumplex diagram. The emotions contained in the emotional corpora used in this study are represented in bold fonts [4]

	Valence yv_e	Arousal ya_e
pleasure	$\cos(\pi/20)$	$\sin(\pi/20)$
happiness/joy	$\cos(3*\pi/20)$	$\sin(3*\pi/20)$
pride/elation	$\cos(5*\pi/20)$	$\sin(5*\pi/20)$
excitement	$\cos(7*\pi/20)$	$\sin(7*\pi/20)$
surprise/interest	$\cos(9*\pi/20)$	$\sin(9*\pi/20)$
anger/irritation	$-\cos(9*\pi/20)$	$\sin(9*\pi/20)$
hate	$-\cos(7*\pi/20)$	$\sin(7*\pi/20)$
contempt	$-\cos(5*\pi/20)$	$\sin(5*\pi/20)$
disgust	$-\cos(3*\pi/20)$	$\sin(3*\pi/20)$
Fear	$-\cos(\pi/20)$	$\sin(\pi/20)$
boredom	-0.5	0
disappointment/frustration	$-\cos(\pi/20)$	$-\sin(\pi/20)$
shame	$-\cos(3*\pi/20)$	$-\sin(3*\pi/20)$
regret	$-\cos(5*\pi/20)$	$-\sin(5*\pi/20)$
guilt	$-\cos(7*\pi/20)$	$-\sin(7*\pi/20)$
sadness	$-\cos(9*\pi/20)$	$-\sin(9*\pi/20)$
compassion	$\cos(9*\pi/20)$	$-\sin(9*\pi/20)$
relief	$\cos(7*\pi/20)$	$-\sin(7*\pi/20)$
admiration	$\cos(5*\pi/20)$	$-\sin(5*\pi/20)$
love	$\cos(3*\pi/20)$	$-\sin(3*\pi/20)$
contentment	$\cos(\pi/20)$	$-\sin(\pi/20)$
neutral	0	0

Figure 5 shows the visual representation of Table 2.

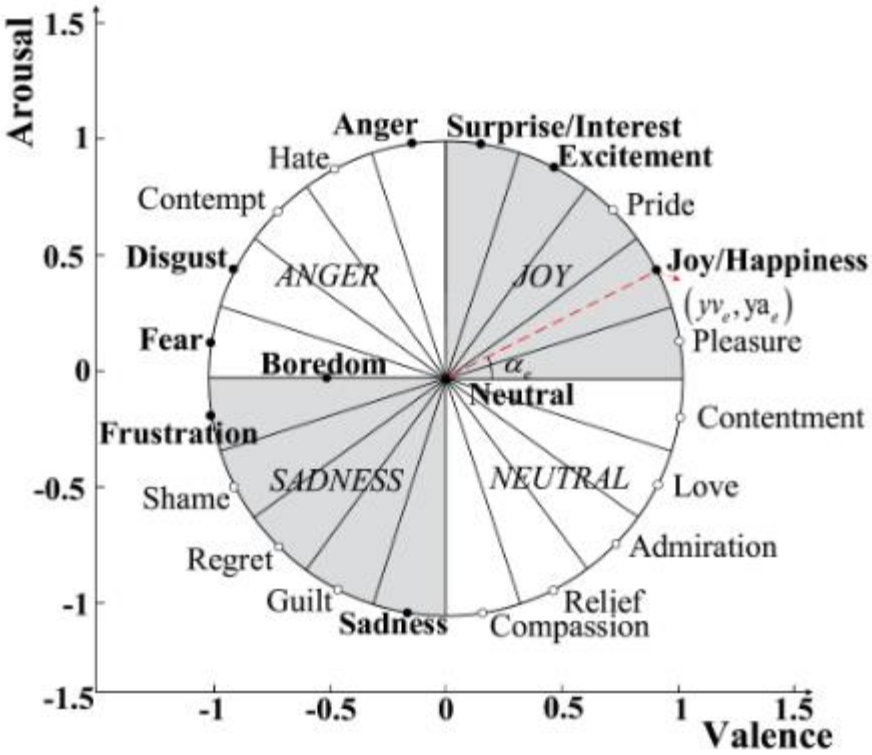

Figure 5 Proposed circumplex diagram of dimensional arousal-valence space. Each emotional state is identified by a point on the unit circle specified by the coordinates (yvi, yai) and the angle αi

The first method is commonly used to create SER models because the Speech Corpora are often labeled with discrete emotions. The second method is commonly used by psychologists to determine emotions. Only a few papers have implemented this method by mapping the discrete emotions from datasets onto the SASM [4], [10]. The selected method narrows the options for each part of the SER model discussed in section 2.3 SER.

2.2.2 Subjectivity in emotion interpretation

Both methods and their possibilities in implementations offer many methods to label emotions. How these emotions are expressed can differ based on language, age, culture, and personality [11]. This means that the emotions in speech corpora could be labeled "subjectively" because they differ on previously mentioned factors [4], [11]. For example, an "introverted" person would express their emotions less intensely than an "extroverted" person. This doesn't mean the emotion is mislabeled, but they could slightly differ from how others perceive them, and their Valence-Arousal values could be shifted from the "true" emotion on the SASM [4]. This is a difficult task to overcome and is attempted in [4]. However, it is also possible to adapt the model for a specific user by determining their "neutral" state to compare it to the average Valence and Arousal [12].

2.3 SER

Several methods for SER have been presented in the state-of-the-art. It is essential to discuss each part of the chain that can be modified to give an overview of how SER works.

In each of these parts, there are multiple options to be selected or combined. This means that there are a significant amount of possible ways to implement SER. Because of this, it is not always possible

to compare one model to another. In sections 2.3.1 until 2.3.7, the state-of-the-art methods will be explored in more detail for each part of SER.

2.3.1 Speech corpus

The speech corpus is the dataset of labeled audio files to perform SER. These are openly available in multiple languages. However, permission was necessary to access the Spanish speech corpora, and the choices were limited. Some commonly used speech corpora are:

- “CASIA is a Chinese corpus collected from 4 Chinese speakers exhibiting 6 emotional states” [13];
- “EMODB is a German corpus that covers 7 emotions by 10 German speakers” [13];
- “EMOVO is an Italian corpus recorded by 6 Italian speakers simulating 7 emotional states” [13];
- “IEMOCAP is an English corpus that covers 4 emotions from 10 American speakers” [13];
- “RAVDESS is an English corpus of 8 emotions by 24 British speakers” [13];
- “SAVEE is an English corpus recorded by 4 British speakers in 7 emotions” [13].

They differ in the number of “actors” or different people present in the samples, the number of samples, the number of labeled emotions, the quality of the samples, and other features. The quality of the samples differs in how the emotions are provoked. The speech corpus could be:

- natural: derived from real-world data;
- elicited: derived from triggered emotions which can be artificial;
- or actor based: derived from trained artists, but can be episodic and very artificial [7].

2.3.2 Conversion of labels

The conversion of labels is only necessary if a transition from discrete to dimensional emotion detection is desired. In this process, the labels in the Speech corpora are transformed into a position on the Valance-Arousal axes, as shown in Figure 5. These labels can then be used as the output of a model. To compare the model's accuracy with the test output, the Euclidean distance between the predicted output and the other values gets measured to find the closest emotion to the predicted emotion to compare the accuracy [4].

2.3.3 Audio preprocessing

Audio preprocessing is the modification of the audio before the features are extracted. This step is only sometimes performed. There are two main types of preprocessing methods. The first is audio framing the second is noise filtering.

Audio framing involves separating audio files into segments of the same length to process them with this fixed length for a fixed amount of features by applying a Hamming window [13].

Noise filtering tries to minimize the unwanted sound that disturbs the speech in the audio [14]. This is often done with a Wiener Filter, Spectral Subtraction, or other methods [7].

2.3.4 Feature extraction

Feature extraction is the part where the audio file will be analyzed in various ways to determine a correlation with the output later. The number of selected features will determine the number of

inputs the model needs to process and will make the model take longer to train. Most of these features are usually extracted using the Python library Librosa. These are some of the most frequently used features:

- MFCC is a human ear-inspired sound frequency representation [15];
- Zero Crossing Rate (ZCR) is the rate of amplitude sign changes in a voice signal [15];
- Spectral Flux is the rate of change in spectral magnitudes [15];
- Chroma is the pitch class energy and quality [15];
- Contrast is the spectral peak and valley differences [15];
- Tonnetz is the representation of harmonic networks in music [15];
- Spectral Centroid is the center of mass in the frequency spectrum [15];
- Spectral Roll Off is the proportion of spectral energy beyond a threshold [15];
- Zero Frequency Filtering (ZFF) is the abrupt closure of the vocal folds [12].

When selecting which features to use, looking at which are more correlated with the expected output is important [16]. In the state-of-the-art, there is a significant difference in which and how many features they select, and in general, it cannot be concluded that more features mean better detection because some papers can achieve state-of-the-art accuracy for discrete emotion estimation while only using MFCC features [13].

2.3.5 Preprocessing of features

After the features are selected, it could be necessary to preprocess the selected features. The first reason is that the number of inputs for the model can be minimized by preprocessing the features. Because of this, the model will then be able to be trained faster [4], [16]. The second reason is that when the features are normalized, the data quality can improve, and the performance of the models can improve [17]. There are three ways to do this. The features can be decreased, sampled, or normalized.

Decreasing the features can be done by removing unnecessary features by leaving them out and comparing performance or by looking at the correlation matrix [4], [16]. If the more correlated features are removed, the performance can be improved, and the model training speed will increase [4], [16].

Sampling the features can be done for some features to decrease the data processed [16]. This can improve the model's performance but will increase the model's training speed [16].

Normalization of features will scale the features to ensure they contribute equally to the model's output [17]. This can improve the model's performance and can improve the model's training speed [17].

2.3.6 Machine learning models

To select the appropriate machine learning model, it is important to consider whether emotions are discrete or dimensional, as this determines whether a classification or regression model should be used. This limits the number of options there are for detection.

In the state-of-the-art, there are many models to choose from to detect emotional speech. The most prominent of these models are Support Vector Machines (SVM), Convolutional Neural Networks (CNN), Random Forest (RF), and Long Short Term Memory (LSTM) models [18]. Combinations exist, such as a model with a CNN input and LSTM hidden layer(s) [18]. The CNN layer would be ideal to extract the right information, and the LSTM layer(s) are ideal to detect patterns in the data [18]. The model choice also depends on what kind of data and how much data will be processed [18].

After selecting a model, finding the right parameters for training is important. This can be done by looking at previous research or experimental research. Some of the parameters that can be selected are:

- Number of Layers is the total number of layers in the model architecture;
- Types of Layers is the different kinds of layers, like convolutional, LSTM, dense, etc;
- Number of Nodes in the Layers is the number of neurons in each layer;
- Types of Activation Functions is the functions that introduce non-linearity in each neuron of a layer, like relu, sigmoid, tanh;
- Number of Epochs is the number of complete training cycles over the dataset;
- Loss Function is the function that measures the model's error during training;
- Optimizer is the algorithm that updates model parameters to minimize the loss or validation loss [19].

The parameters chosen to train the model also determine the model's performance on the data. CNN and LSTM are deep learning models, meaning they can improve when more data is provided, in contrast to SVM and RF which are machine learning models and can benefit from more data until a certain point [18]. However, the quality of the data and how it is preprocessed will also influence the model's performance [4].

2.3.7 Evaluation of model

The type of emotions must be considered to evaluate the model as they can be discrete or dimensional. This changes our outputs to be respectively classification or regression. Classification can be easily defined by looking at the accuracy or F1-score of the test emotions compared to the predicted emotions. The regression is more difficult. This works by looking at how close the test outputs are to predicted outputs using the Euclidean distance, this can be measured with the concordance correlation coefficient (CCC) [4]. The result can be seen visually by plotting both results on a graph and observing how close they are to the actual value. And later the accuracy and F1-score can defined by converting the labels from the valence and arousal to the discrete emotions.

Furthermore, the model's performance can be assessed using a confusion matrix to identify the misidentified emotions. It is also important to check if the model is overfitting on the speech corpus data. This will result in lower performance on new data.

2.4 State-of-the-art results of SER

In the state-of-the-art SER is performed a lot with classification, so with discrete emotions. These often get results above 80 and 90% [7]–[9]. However, SER with dimensional values is less prominent. This is probably because discrete emotions are easier to predict for people than Valence-Arousal values because it harder to annotate [4]. Aside from this, "emotion corpora with dimensional emotion labels are rare"[4].

There is limited research to be found about SER with dimensional values and of the research that exists, they sometimes use a rSASM instead of SASM or they use it to detect with less emotions mapped on the axes. This could increase the simplicity because fewer possible emotions can be detected incorrectly.

2.5 Considerations for real-world applications of SER

In the real world, many considerations have to be made for machine learning models.

The first consideration must be that the models can perform better if optimized for certain languages, cultures, personality types/traits, genders, or age ranges just because they know which model to select for this characteristic [10], [20], [21]. This is hard to achieve because there are few speech corpora with all of this data to train on.

The second consideration is that noise, pets, or multiple people in the audio recording could influence these results [10], [14], [22]. Because also a TV that is turned on, a phone call in the background, or noise from outside could be detected by our model and could lead to incorrect results [10], [14], [22].

The third consideration is that emotions can also be specific to sleep quality, the context of the conversation, specific expressions, weather, situational awareness, and other things that are not considered for now. This could also influence how the emotions are expressed and thus also our results.

The last consideration is that audio is only a part of a person's emotion. That's why the ADDIM project also has integrated the context of all the sensors and visual emotion estimation. This SER model will be part of the emotion and mood estimation of the ADDIM system.

2.6 Related work on ER with robots and elderly care

The HIMTAE project [23] also aims to enhance the user's quality of life by using a robot and detecting their emotions using the sensors. They also recommend including SER.

In another study [24], they look at the effects of the robot named Ryan with or without emotional analysis in a conversation with the user. And they noticed that there was not much difference in the empathic or non-empathic version of the robot in conversation with the user. However, when the word count was measured, and the exit survey of all participants was finished, they noticed that the emotional version of Ryan was more engaging and likable by the participants. And that the empathic version of the robot encourages users to have longer conversations compared to the non-empathic version. They suggest that this has the potential to decrease depression in the users.

Both projects suggest a quality of life improvement when emotion detection is implemented. Speech emotion recognition could be a crucial part of mood estimation [23].

3 Implementation & integration

This section will focus on how the ADDIM system works and how the SER model was integrated into the system.

Section 3.1 describes the implementation details of the current ADDIM system.

Section 3.2 describes the integration of SER into the ADDIM system.

3.1 Implementation of the ADDIM system

Figure 6 shows the overview of the current ADDIM system. This is still a work in progress and thus will change later on. For example, the testing robot will be changed to a commercial robot from INTEC Sistemas.

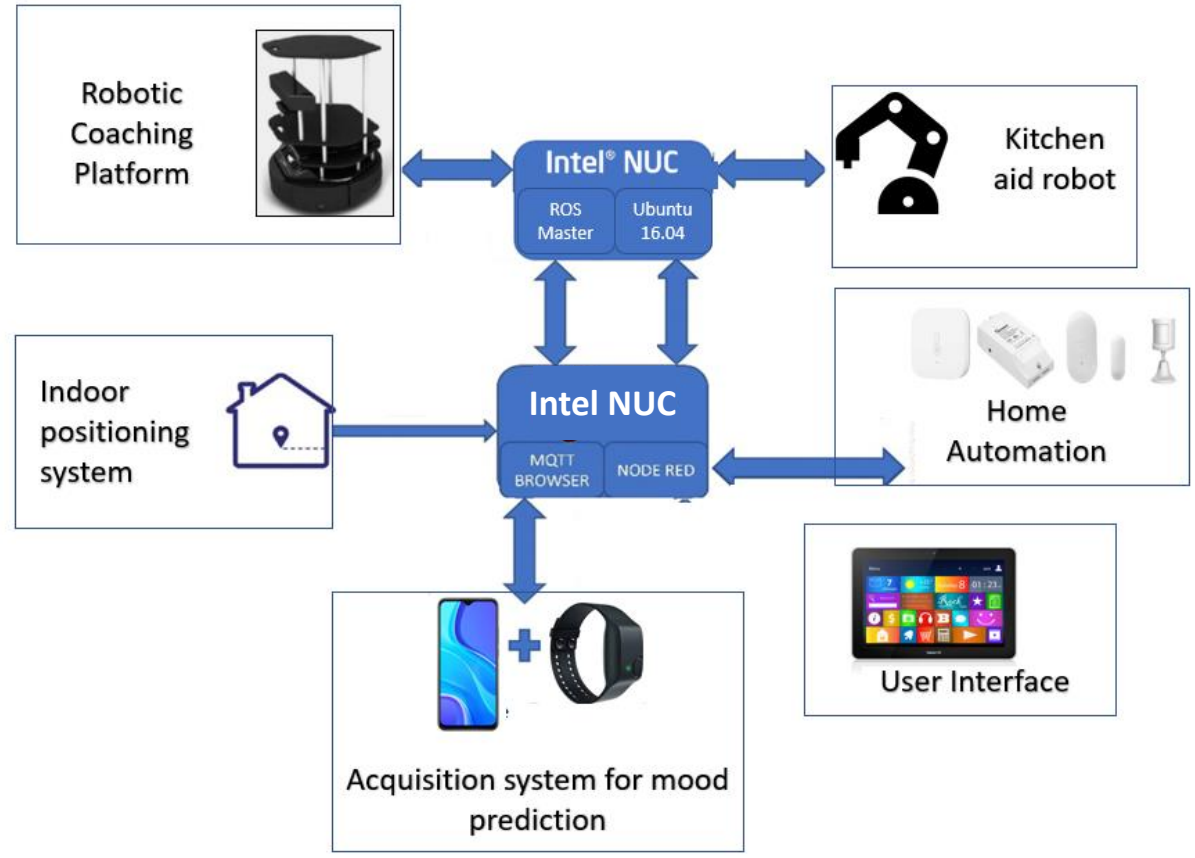

Figure 6 the ADDIM system with the experimental robot

The ADDIM system contains three main parts, the robot, the sensors and the Intel NUC as the central part of the system:

- The robot is now an experimental version which is made with a 3D printed enclosure. It consists of a Turtlebot II base for movement. The rest of the robot is mounted on top of this with a Kobuki's IClebo mobile base. This keeps the components together, such as the Intel NUC, the Alexa Echo in its head, and the LIDAR Hokuyo UST-10LX. Which are the brains of the robot and the sensors to detect obstacles in the environment and create a map. The new robot will be a commercial robot provided by INTEC Sistemas.
- The sensors are Internet-of-Things (IoT) and self-made sensors communicating over Zigbee or Wi-Fi. These detect if the user is in the bed, or chair, opening a door, using the sink, and more to know the user's activities.

- The Intel NUC is the central part of the system because it connects everything together and decides what happens in the system. It runs Ubuntu 16.04 with Docker containers. In these Docker containers is our modular system which works together to make the system easily expandable. These are the docker containers:
    - Mosquitto is the docker container that runs the Message Queuing Telemetry Transport (MQTT) Broker. This means that it is responsible for all MQTT communication between some devices and services. The devices can publish information on topics or subtopics and broadcast them to all other devices subscribed to them.
    - Zigbee2MQTT is the docker container that uses a Zigbee antenna to communicate with Zigbee devices to connect them to the MQTT Broker.
    - MariaDB is the docker container that stores the Home Assistant data and user detection data of the SER model.
    - Influx DB is the docker container used to store the Home Assistant data.
    - Gunicorn is the docker container used to communicate with the Alexa Echo in the robot.
    - Nginx is the docker container used for the Gunicorn connection to the Amazon Servers.
    - NodeRed is the docker container used to automate the system and is integrated into Home Assistant with a plug-in.
    - Home Assistant is the docker container that connects all parts together and provides a graphical interface to visualize the state of sensors and parts of the system.

The decision-making was done in NodeRed by predefining schedules and actions to be performed. These predefined schedules and actions are connected to the whole system and can read the inputs from the sensors and robot. From these inputs, the next steps are defined. And will be sent through the system and the robot to be performed.

NodeRed was chosen because of its ease of integration with the rest of the system and the visualized way of automating it. However, it is operating slower, making it less reliable.

3.2 Integration into the ADDIM system

To integrate SER into the ADDIM system, an extra docker container was added to the system. The docker container is a Python container that will train and run the model. And because the NodeRed was unreliable it was a good idea to remove the NodeRed container and translate the automatizations using Python. This will make it easier for later improvements of the ADDIM system to add a mood estimation model based on all the sensors and other models' outputs.

The Python container communicates with the sensors through:

- MQTT;
- WebSocket for the robot;
- HTTP requests to Home Assistant for Alexa.

The last two will change with the commercial robot. This is possible because the Python code is written in modules that can easily be changed.

To perform SER in this system, the automatization will trigger Alexa to ask the user a question predefined by a team of psychologists. The response will be recorded in the Intel NUC in the robot using a Microphone attached to it. The audio file is sent back to the automation code. It is then given as an input to the SER model and the output is stored in MariaDB. Afterward, the automation code

can handle the response with an adequate action if necessary by combining all the information from all the inputs. The robot could for example suggest calling the user's family in case the user feels sad.

Simultaneously, the audio file is stored in a separate folder for the selected user for retraining when the system is less utilized. This way, the model will be better adapted to the specified user over time. The estimated emotion from the ADDIM system can be used to confirm that the prediction of SER was correct. With this data, SER can be retrained with the user's data and become more accurate over time.

4 Results and discussion

This chapter consists of the explorative part, and the following chapters will explore the fittest solution to be implemented and explain the approach taken to achieve these results.

Section 4.1 describes which emotion interpretation approach was taken.

Section 4.2 describes which speech corpora were taken.

Section 4.3 describes the audio processing that was performed.

Section 4.4 describes features that were extracted and what libraries were used.

Section 4.5 describes how the features were preprocessed.

Section 4.6 describes how the machine learning models were evaluated.

Section 4.7 describes how the machine learning models performed.

Section 4.8 summarizes the whole approach that was taken to perform SER.

4.1 Interpreting emotions

The first tests were performed with discrete emotion interpretation, however in the ADDIM system, the dimensional values are necessary to provide a granular approach in emotion detection to train a Machine Learning algorithm to determine the emotion based on all the other parameters. Using the output of this multimodal model, the mood can be estimated. Because of this, the next choices will be adapted to the optimization of detecting dimensional values. These were more difficult to predict compared to discrete emotions. Juan Pedro Serna, a physician and CEO of E-Doctor, advised this approach.

4.2 Speech corpus

The speech corpora that were selected are EmoDB, EmoFilm, and Ravdess. The emotion categories were balanced after normalization. EmoDB and Ravdess were selected because of their frequent usage to perform SER. EmoFilm was selected because it was the only available speech corpus with Spanish audio files.

These speech corpora were combined to extend the available emotions and data samples to generalize the model and avoid overfitting one speech corpus. It is important to combine these datasets because the balanced dataset removes a significant amount to avoid biases.

4.3 Audio preprocessing

There is no audio preprocessing performed. This was unnecessary because the user is assumed always to answer and a constant noise level is assumed. However, the audio file of the user is saved for retraining.

4.4 Feature extraction

The extracted features are a combination of the mean, minimum, maximum, or standard deviation of the following feature sets:

- Short-Time Fourier Transform (STFT) Magnitude is the frequency domain representation over time intervals, revealing signal components;
- Pitch and Magnitude Tracking are the trace changes in pitch and magnitude of the audio signal;
- Pitch Statistics are the statistical properties of pitch variations in the signal;
- Spectral Centroid is the center of mass in the frequency spectrum;
- Spectral Flatness measures how flat or peaky the spectrum is;
- Mel-Frequency Cepstral Coefficients (MFCCs) is a human ear-inspired sound frequency representation;
- Chroma is the pitch class energy and quality;
- Mel Spectrogram shows the intensity of frequencies over time;
- Spectral Contrast measures the difference between peak and valley in spectrum;
- Zero Crossing Rate is the rate of amplitude sign changes in a voice signal;
- Root Mean Square (RMS) Energy is the average energy magnitude of the signal.

This amounts to 320 features, these are then combined with the output into a correlation matrix to visualize their correlation towards the output parameters. The chosen output parameters are Valence and Arousal. Dominance was not considered as the datasets do not have labels to detect this parameter and because “Valence and Arousal account for most of the independent variance in affective responses” [11].

These features are chosen after extensive testing and selecting the best features to get the most accurate output.

4.5 Preprocessing of features

The features were all normalized and afterward decreased by their relevancy to the output features. The features were decreased using Principle Component Analysis (PCA), this reduces the output features to the requested amount of features while also making sure the features are as less as possible correlated with each other.

4.6 Evaluation of model

To evaluate the model, the following parameters were chosen:

- Accuracy;
- Mean Squared Error (MSE);
- Root Mean Squared Error (RMSE);
- R-squared;
- CCCV;
- CCCA.

However, the accuracy compares the models to select the best-performing features and model parameters. The models were trained on the validation set

Aside from this, the predicted values are plotted to see the quality of the features and the model because the model could have a poor performance, but relatively good accuracy, which was not useful if only two of the seven emotions were correctly predicted.

4.7 Machine learning models

The CNN, LSTM, and a combination of both were chosen from all the possible models. This is because they can perform regression, but are also deep learning models which will have an advantage for retraining later on. First the model description will be briefly explained, afterwards, the training results will be discussed for all the datasets combined which are Ravdess, EmoFilm and EmoDB, compared with the EmoDB results. EmoDB is chosen as reference point to compare to similar research which also used EmoDB [4].

4.7.1 Model description

Figure 7 shows the CNN model used, each block represents a layer with the first one being the input layer. At the left of each block the layer type and activation function are mentioned (if present). At the left of each block the input and output parameters are mentioned.

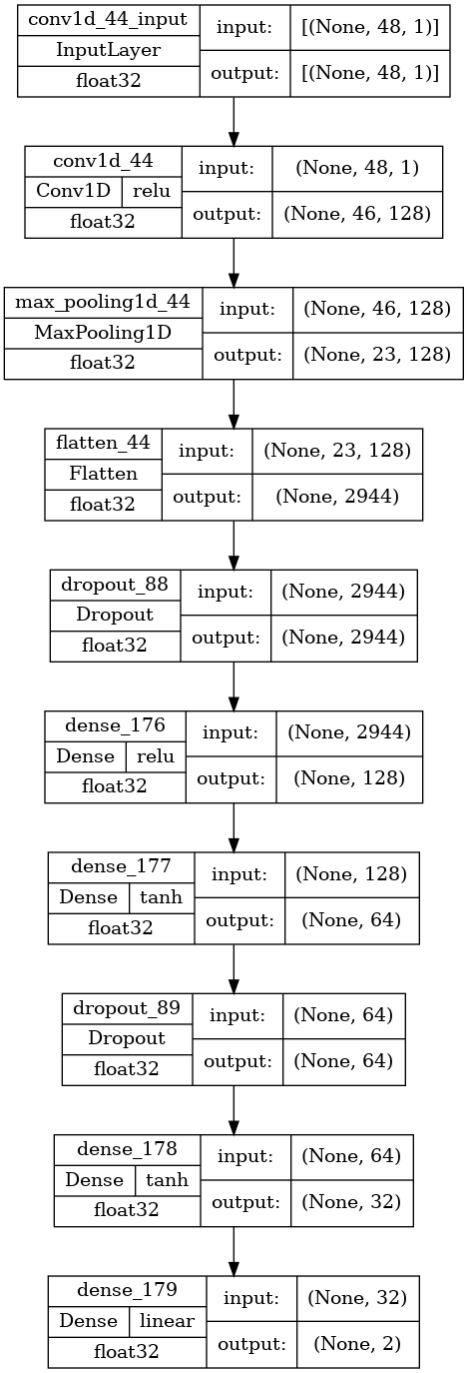

Figure 7 CNN model layer details

Figure 8 shows the LSTM model used, each block represents a layer with the first one being the input layer. At the left of each block the layer type and activation function are mentioned (if present). At the left of each block the input and output parameters are mentioned.

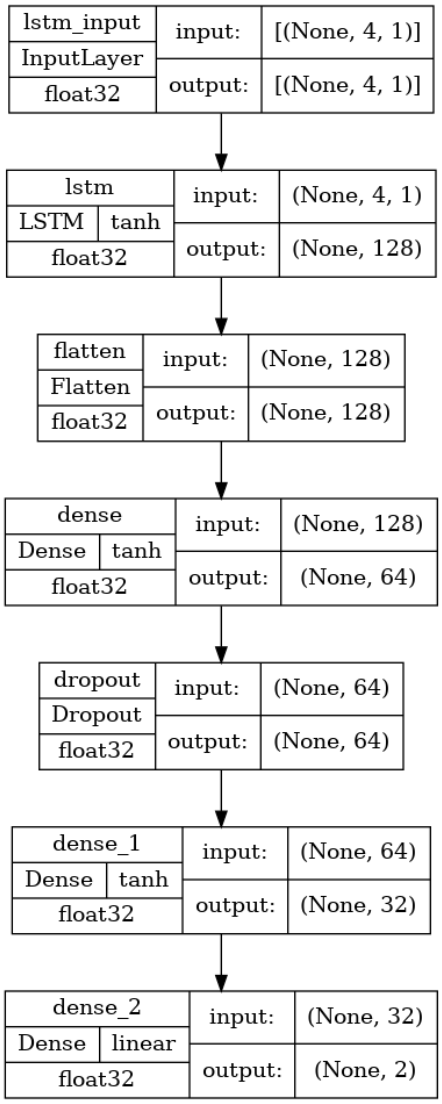

Figure 8 LSTM model layer details

Figure 9 shows the CNN+LSTM model used, each block represents a layer with the first one being the input layer. At the left of each block the layer type and activation function are mentioned (if present). At the left of each block the input and output parameters are mentioned. The first block is the CNN layer and the last block is the LSTM layer in this model.

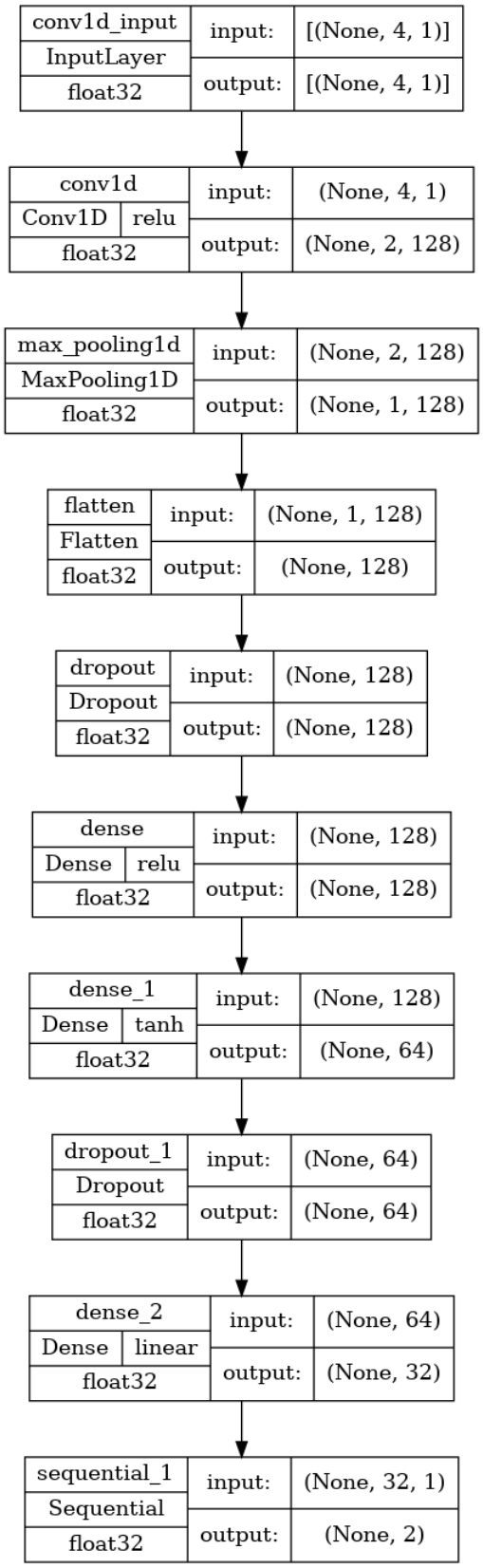

Figure 9 CNN+LSTM model layer details

All the models were trained with the 'Huber Loss' loss-function as it was shown to perform the best during the experiments. For the optimizer, Nadam was chosen with a learning rate of 0.001 for the same reason. The hyperparameters were tuned with the validation set using an Earlystopping function to find the best possible parameters.

4.7.2 Model training results

Table 3 shows the training results for the three models and the difference of using three datasets compared to one. The accuracy of all the datasets combined is significantly lower compared to discrete values, this was expected because of the higher amount of emotions that can be misinterpreted. However, the CNN+LSTM for all datasets CCCV is 7% lower and CCCA is 17% higher compared to the research done in [4]. This is a satisfactory results as this shows that even with 3 datasets the results can be close or better than the state-of-the-art with all emotions on the SASM. The results for EmoDB are even more impressive, the CNN model and CNN+LSTM model achieve the same accuracy, however, the CNN+LSTM model achieves this result with 9 less features. The CCCV and CCCA values of CNN and CNN+LSTM both significantly higher compared to the results from [4].

Table 3 model training results for CNN, LSTM and CNN+LSTM on all speech corpora, compared with EmoDB

	CNN		LSTM		CNN+LSTM	
Datasets	All	EmoDB	All	EmoDB	All	EmoDB
Accuracy	44.94 %	78.78 %	30.33 %	36.36 %	46.06 %	78.78 %
MSE	0.3462	0.1110	0.3357	0.2501	0.3302	0.1269
RMSE	0.5884	0.3333	0.5749	0.5001	0.5746	0.3562
R^2	0.0550	0.6553	0.1335	0.2429	0.1158	0.6413
CCCV	0.4790	0.8531	0.1706	0.5573	0.4557	0.7615
CCCA	0.4952	0.8331	0.4101	0.5253	0.5144	0.8901
Inputs	48	46	6	41	59	40

Table 4 shows the dataset size for the training, validation and testing. These are lower compared to the original dataset because the dataset was not balanced.

Table 4 dataset size after balancing and train-valid-test split

	All	EmoDB
Training set size	623	225
Validation set size	179	64
Testing set size	89	33

Figure 10, Figure 11 and Figure 12 show the Valence-Arousal plot of the CNN, LSTM and CNN+LSTM model trained on all the datasets respectively. In the plot is the predicted value shown as a bullet, the test value shown as a '+' and the incorrectly categorized values are crossed with a red cross.

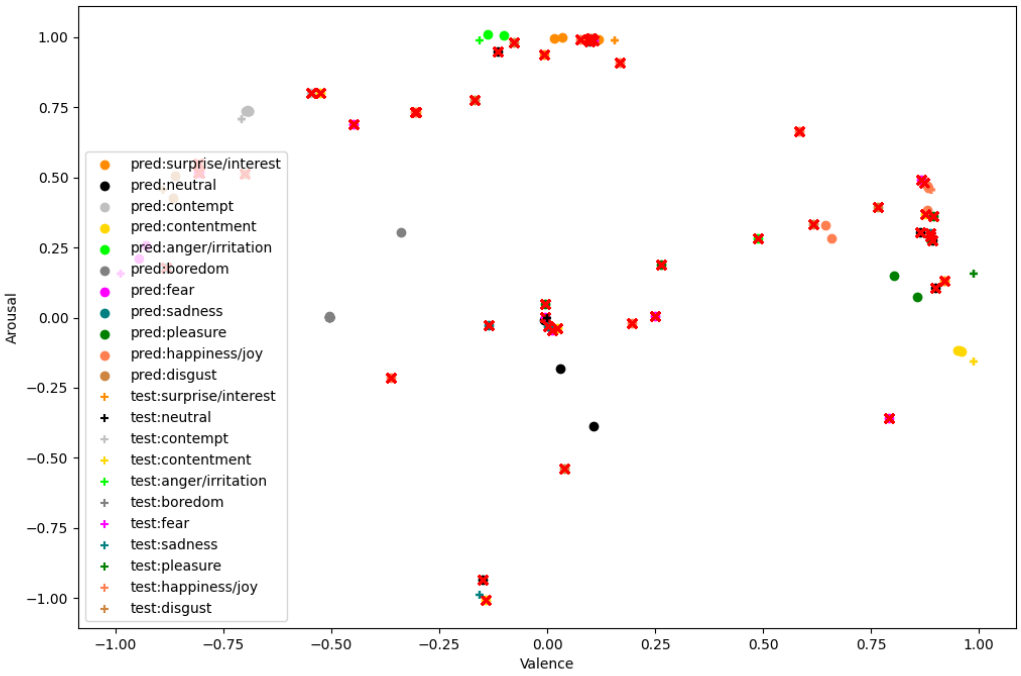

Figure 10 Valence-Arousal plot of the CNN model for all datasets

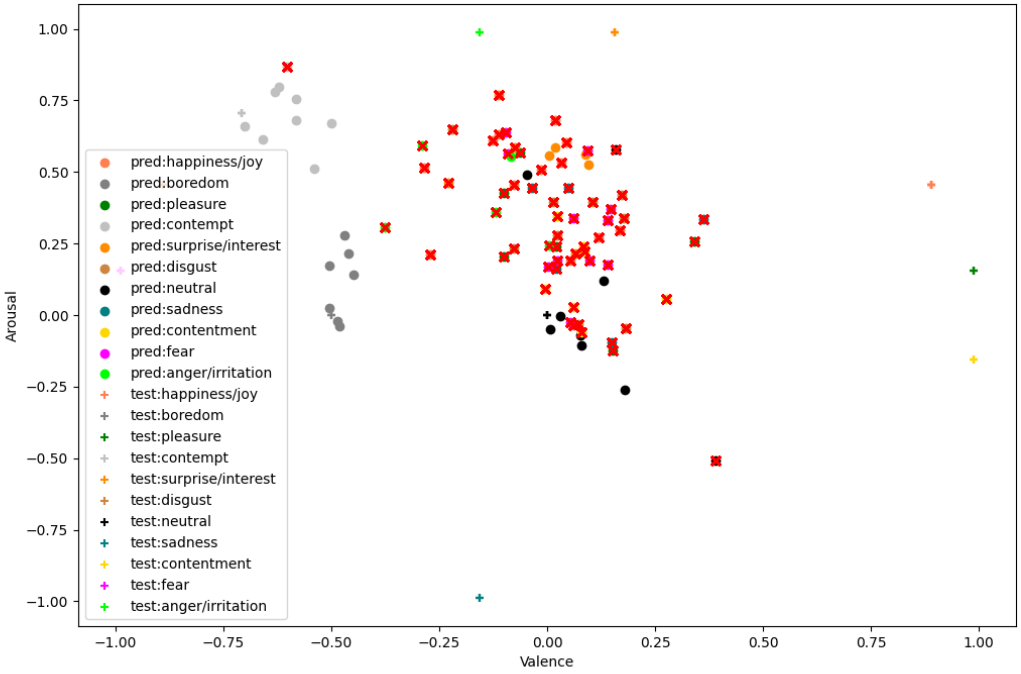

Figure 11 Valence-Arousal plot of the LSTM model for all datasets

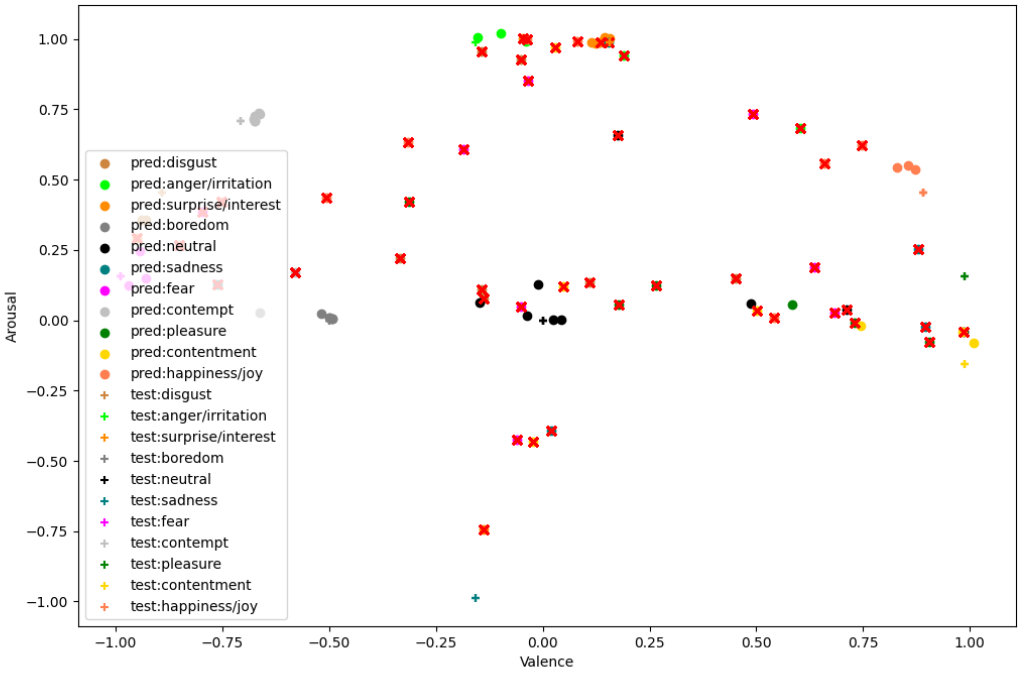

Figure 12 Valence-Arousal plot of the CNN+LSTM model for all datasets

Figure 13,Figure 14 and Figure 15 show the Valence-Arousal plot of the CNN, LSTM and CNN+LSTM model trained on the EmoDB dataset respectively. In the plot is the predicted value shown as a bullet, the test value shown as a '+' and the incorrectly categorized values are crossed with a red cross.

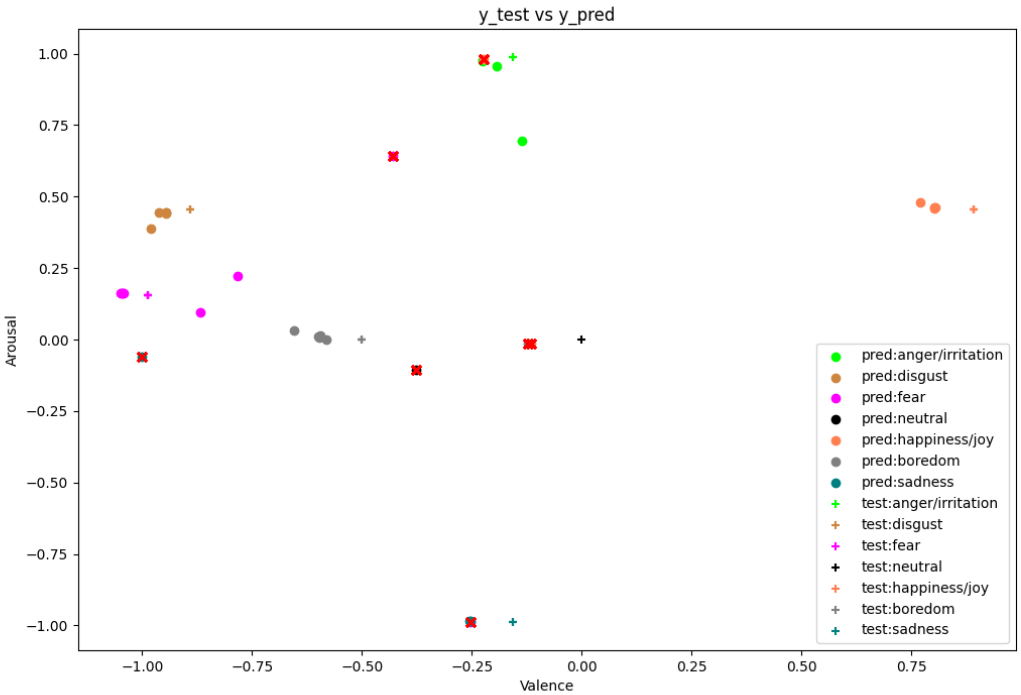

Figure 13 Valence-Arousal plot of the CNN model for the EmoDB dataset

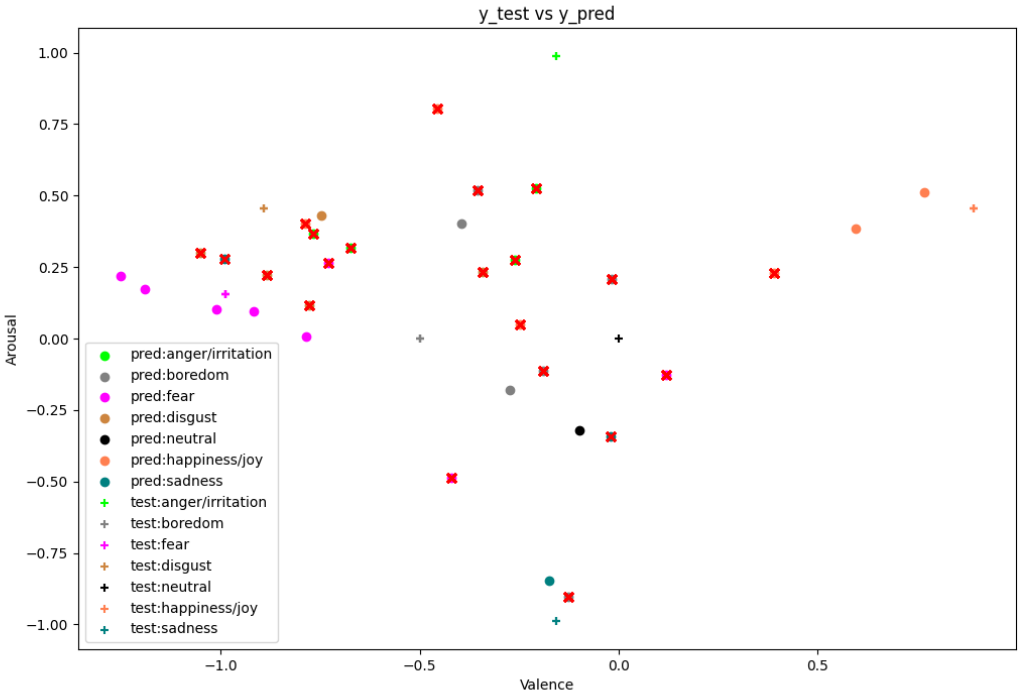

Figure 14 Valence-Arousal plot of the LSTM model for the EmoDB dataset

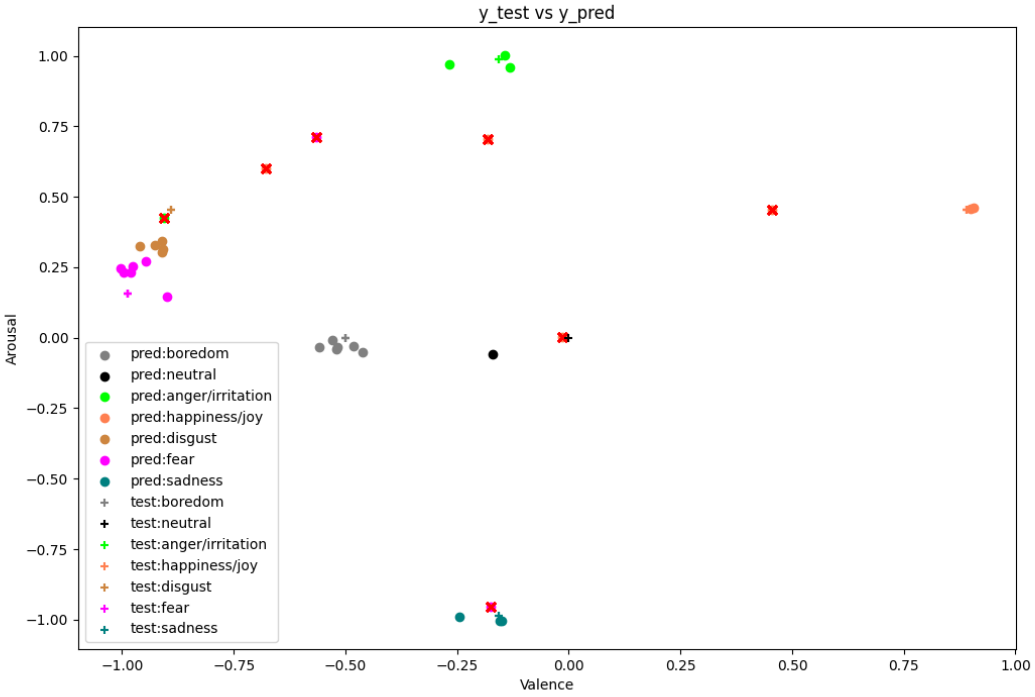

Figure 15 Valence-Arousal plot of the CNN+LSTM model for the EmoDB dataset

Figure 16,Figure 17 and Figure 18 show the confusion matrix of the output of the CNN, LSTM and CNN+LSTM model trained on all the datasets respectively.

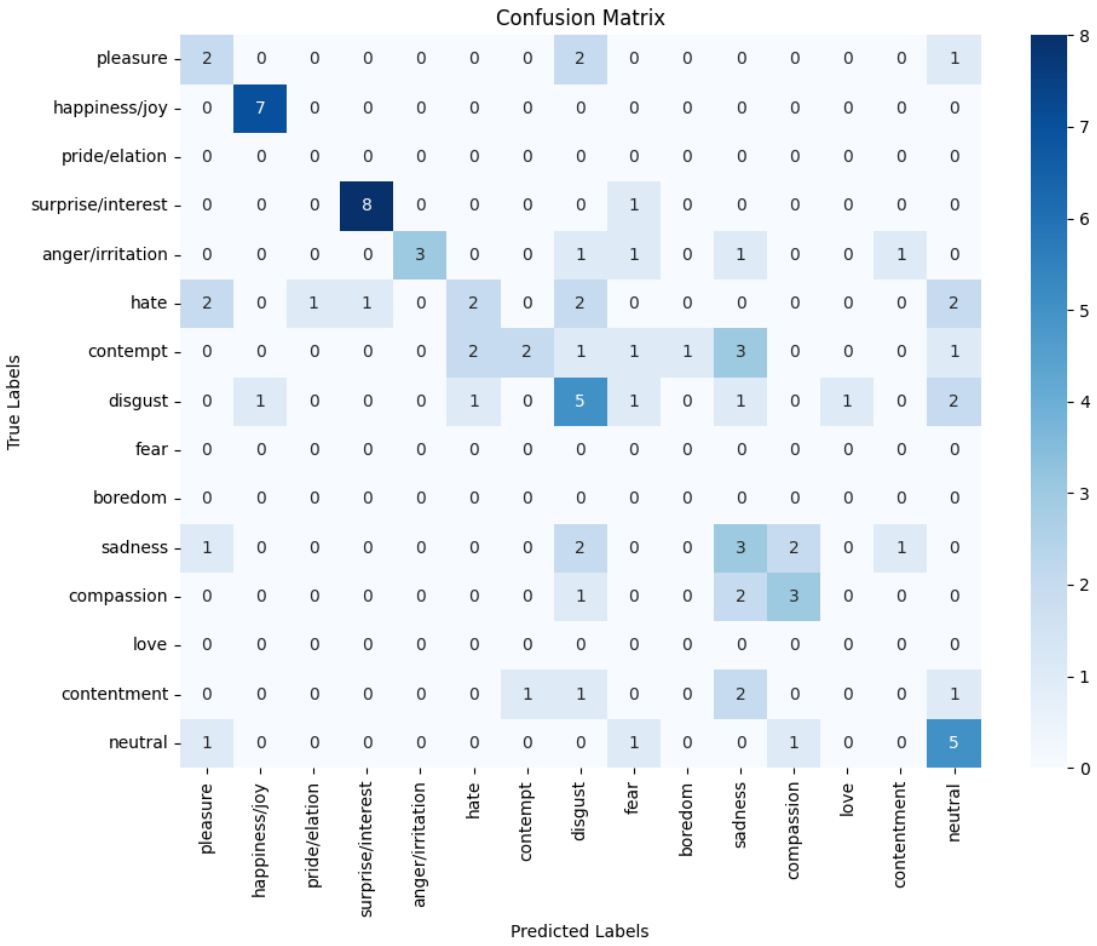

Figure 16 confusion matrix CNN model with all datasets

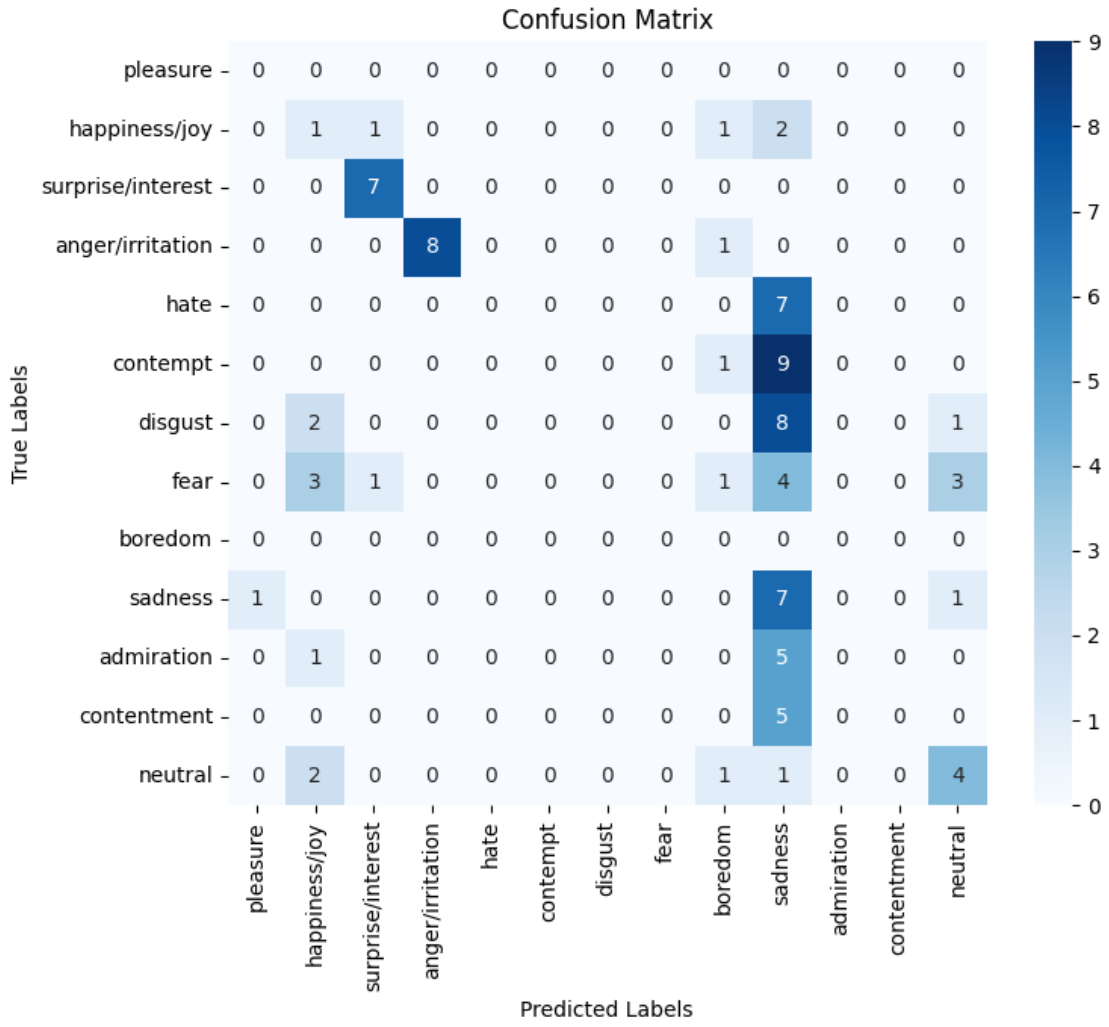

Figure 17 confusion matrix LSTM model with all datasets

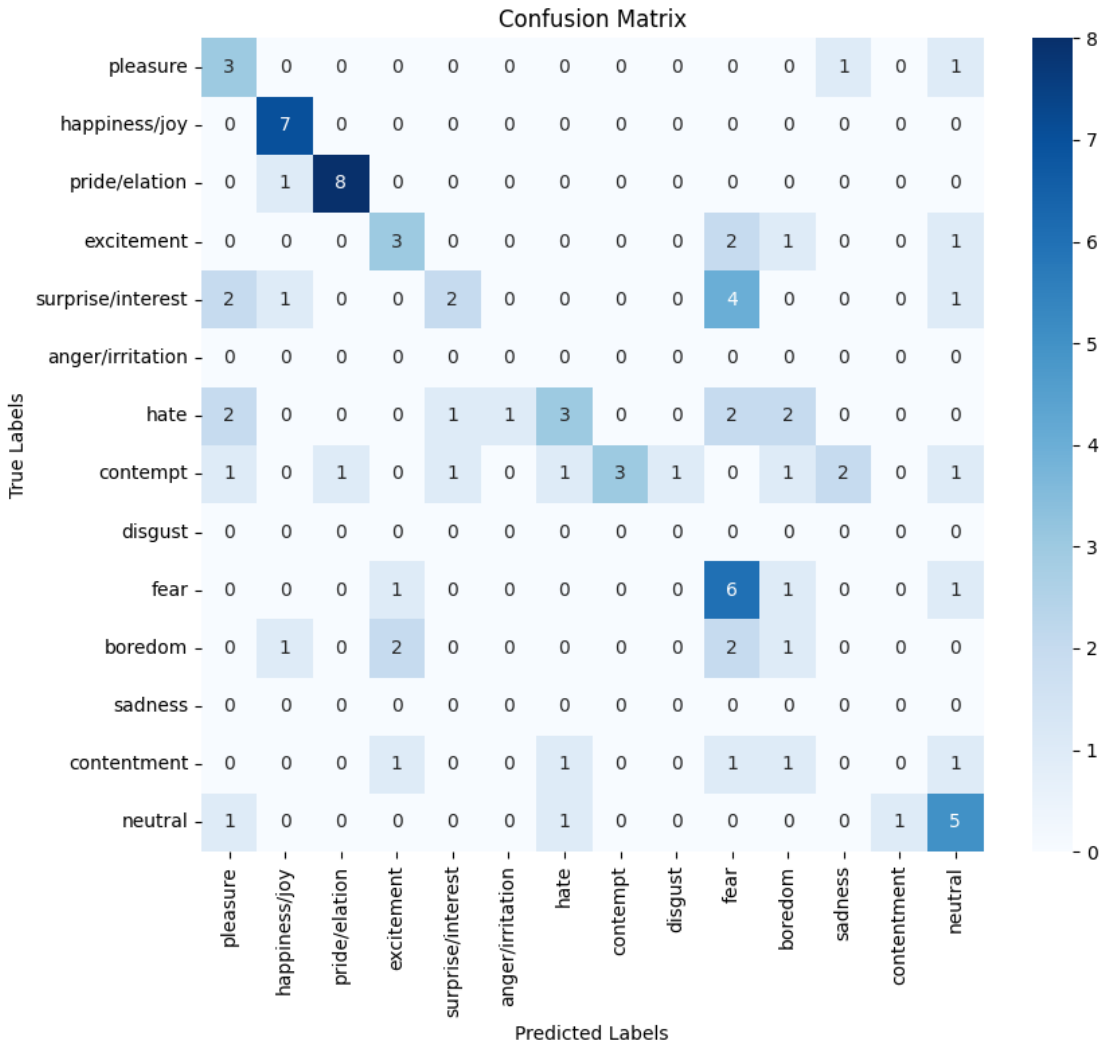

Figure 18 confusion matrix CNN+LSTM model with all datasets

Figure 19, Figure 20 and Figure 21 show the confusion matrix of the output of the CNN, LSTM and CNN+LSTM model trained on the EmoDB dataset respectively.

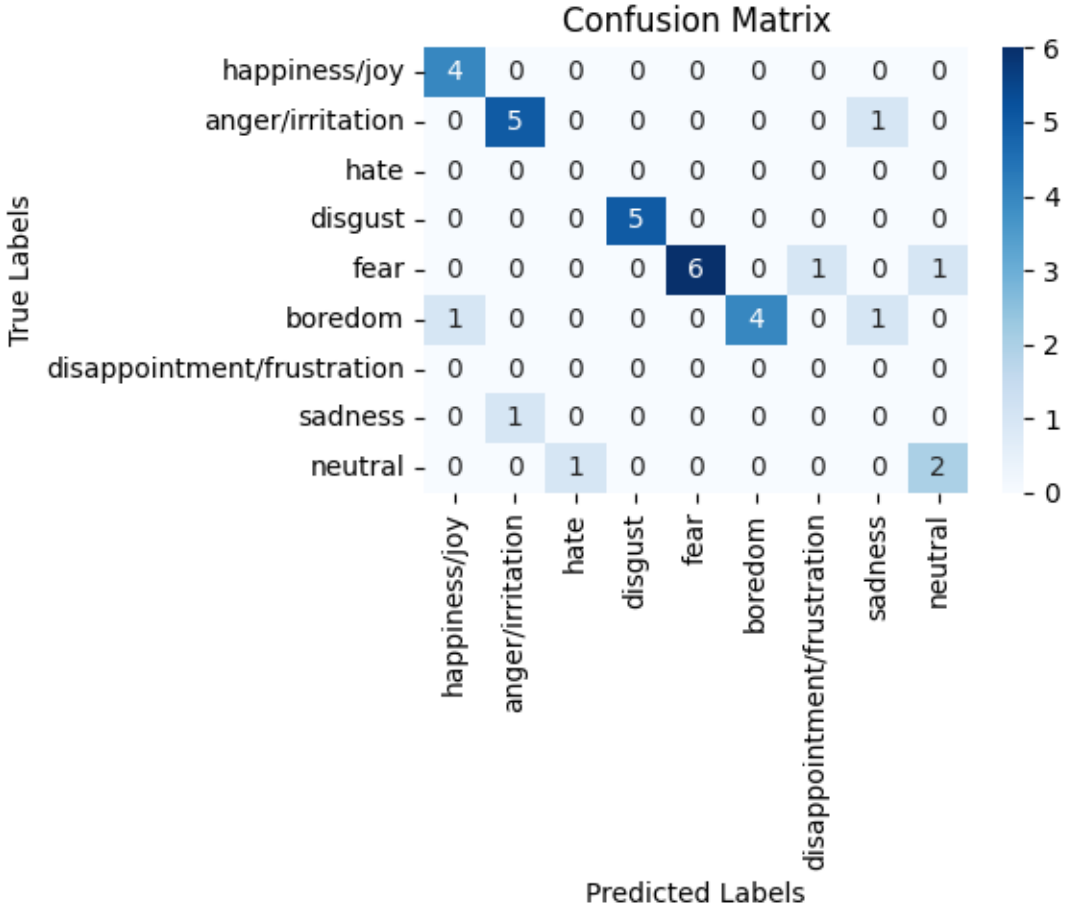

Figure 19 confusion matrix CNN model with the EmoDB dataset

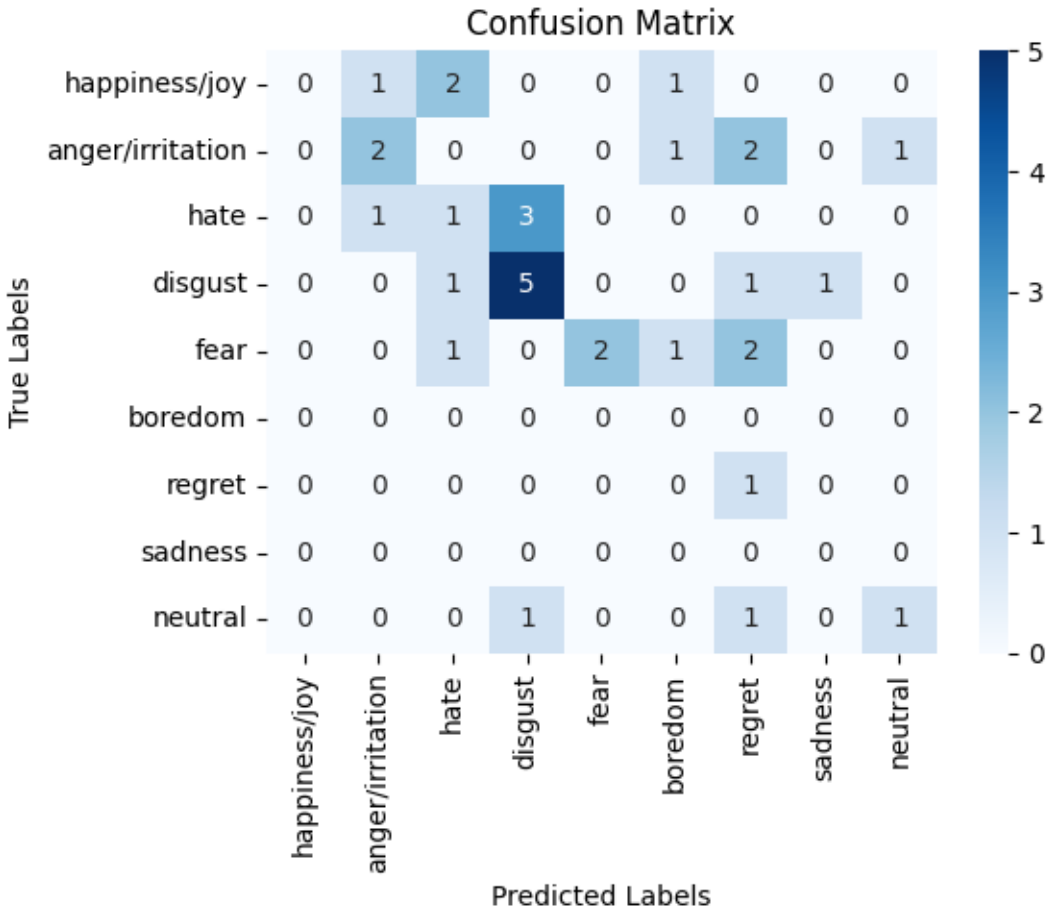

Figure 20 confusion matrix LSTM model with the EmoDB dataset

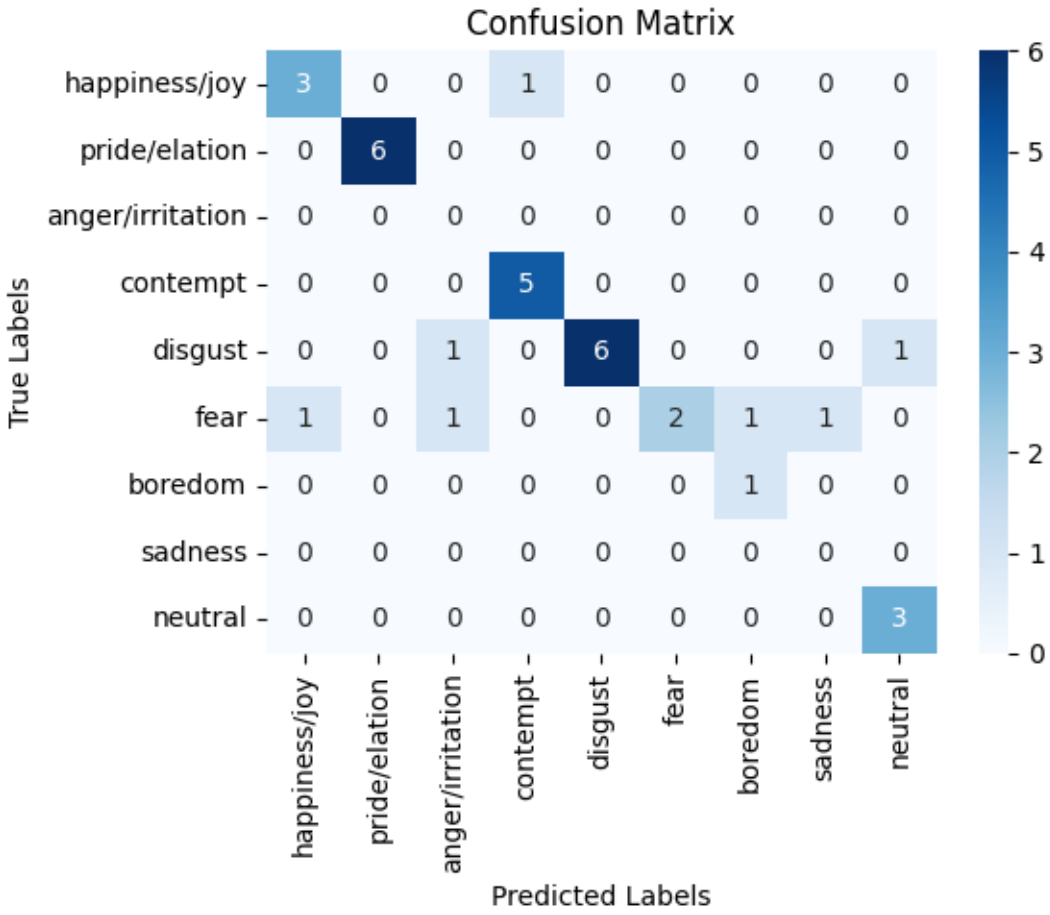

Figure 21 confusion matrix CNN+LSTM model with the EmoDB dataset

Figure 22, Figure 23 and Figure 24 show the training and validation loss plot of the CNN, LSTM and CNN+LSTM model trained on the EmoDB dataset respectively.

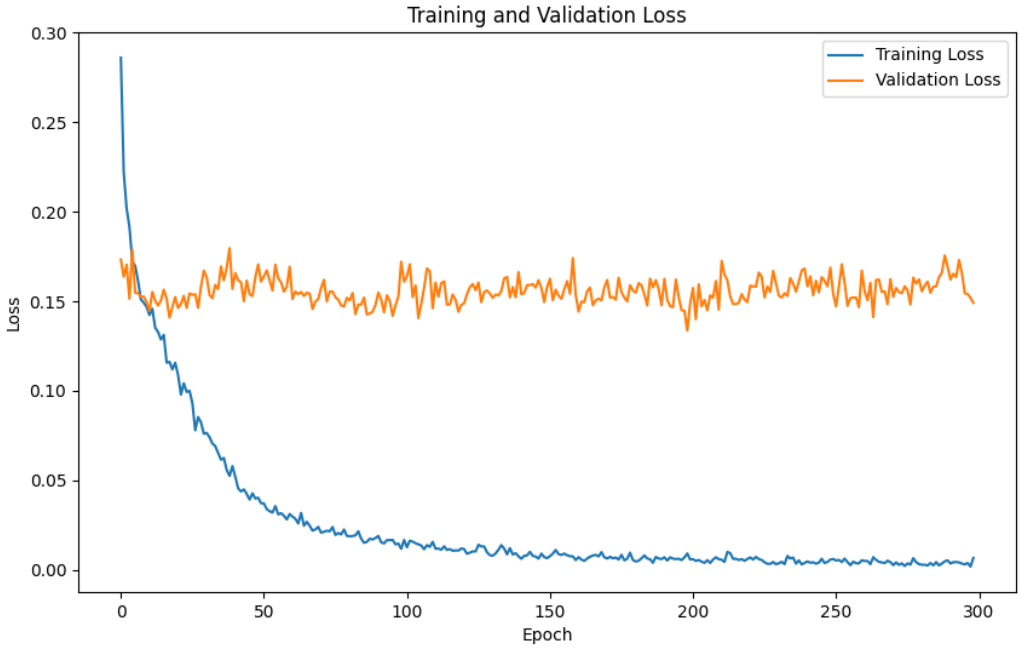

Figure 22 training and validation loss during training of the CNN model on all datasets

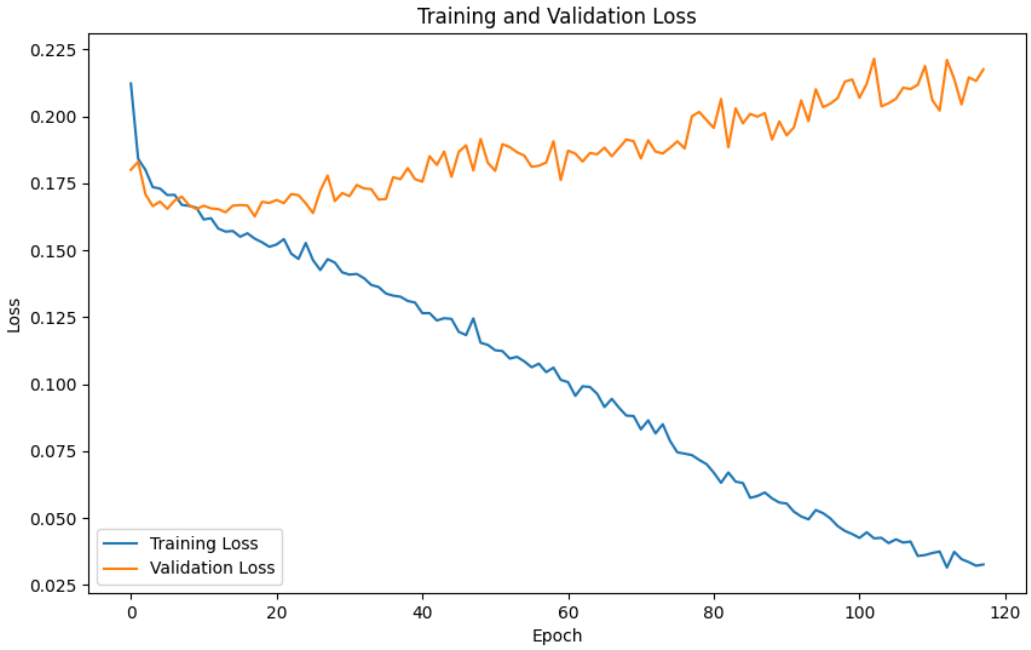

Figure 23 training and validation loss during training of the LSTM model on all dataset

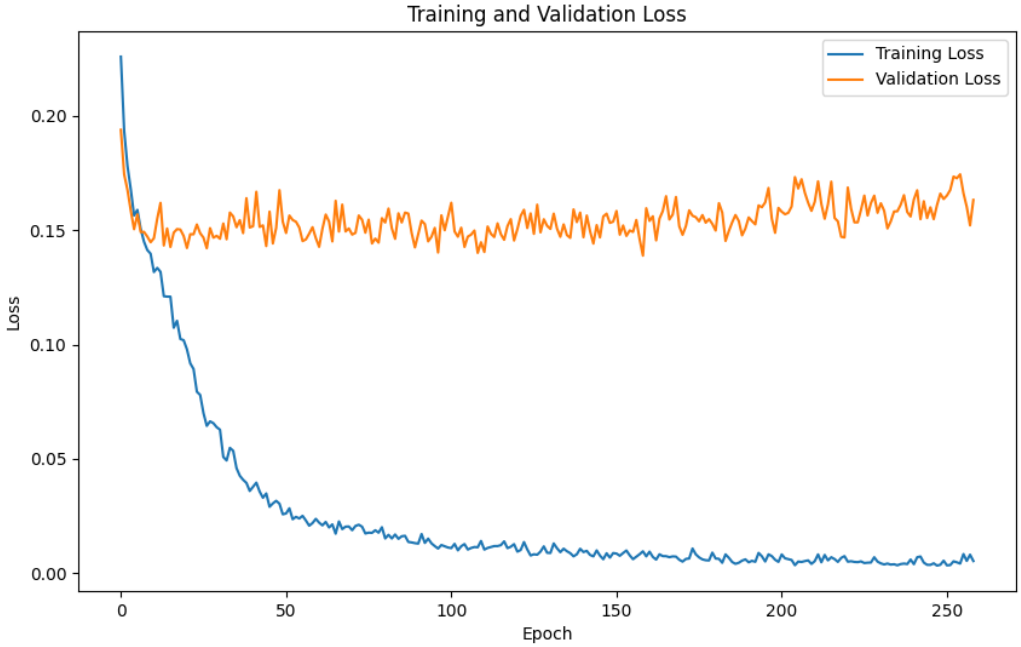

Figure 24 training and validation loss during training of the CNN+LSTM model on all datasets

Figure 25,Figure 26 and Figure 27 show the training and validation loss plot of the CNN, LSTM and CNN+LSTM model trained on the EmoDB dataset respectively.

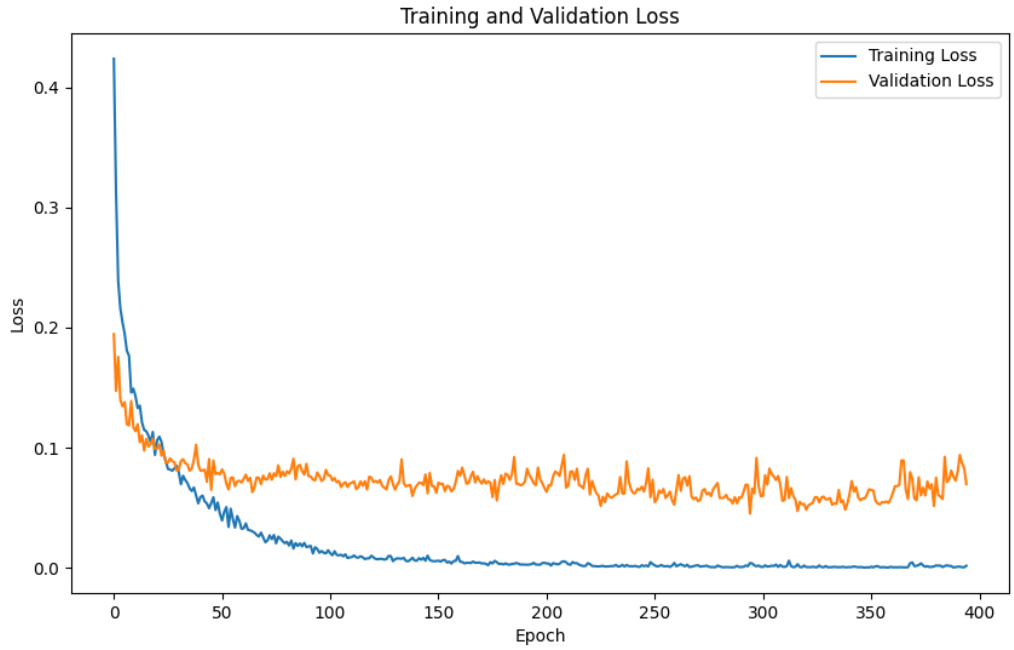

Figure 25 training and validation loss during training of the CNN model on the EmoDB datase

Figure 26 training and validation loss during training of the LSTM model on the EmoDB dataset

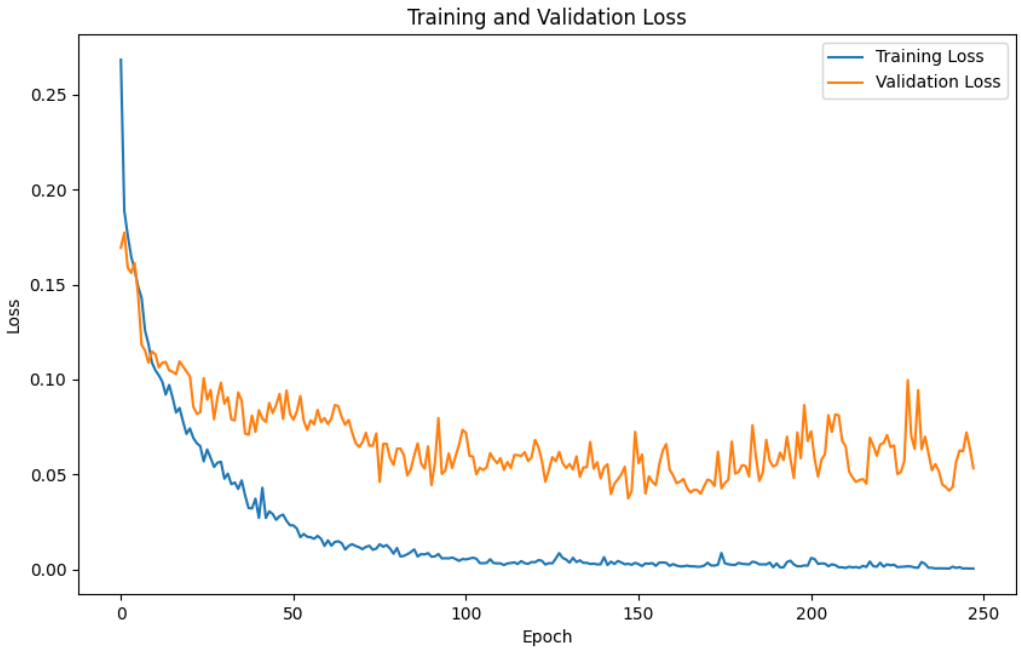

Figure 27 training and validation loss during training of the CNN+LSTM model on the EmoDB dataset

From the Valence-Arousal plots and the training and validation loss plots it can be concluded that the CNN and CNN+LSTM perform similar. However, from the CNN+LSTM seems to slightly outperform the CNN model when comparing the details from Table 3.

4.8 Considerations for Real-World Application of SER

To use the model in the real world, some limitations have to be considered. The datasets used are not representative with the real world. However, by combining them and not overfitting on them, the results could improve over time with more data and retraining to the needs of the user. In future work, this can be evaluated and maybe be improved by adding audio preprocessing. Future work could also check the importance of the Dominance axis or if the model could be adapted to work in real-time.

5 Conclusion

The CNN and especially the CNN+LSTM model are most fit for the ADDIM system as shown in the Results and Discussion. This because it can adapt well to the Valence-Arousal axes on the EmoDB dataset, but also doesn't seem overfit on all the datasets together or the EmoDB dataset. The models are able to get around 50% accuracy on all the datasets together, which is in line with the state-of-the-art Valence-Arousal SER with all the emotions of the SASM.

However, the models still have to be proven to work in the real world in future work.

References

[1] "Spain data | World Health Organization." https://data.who.int/countries/724 (accessed Aug. 18, 2023).

[2] "Ageing Europe - statistics on population developments." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_population_developments (accessed Aug. 18, 2023).

[3] "Suicide rates." https://www.who.int/data/gho/data/themes/mental-health/suicide-rates (accessed Aug. 18, 2023).

[4] S. Jing, X. Mao, and L. Chen, "Automatic speech discrete labels to dimensional emotional values conversion method," *IET Biom.*, vol. 8, no. 2, pp. 168–176, 2019, doi: 10.1049/iet-bmt.2018.5016.

[5] C. Kaufmann, N. Agalawatta, E. Bell, and G. S Malhi, "Getting emotional about affect and mood," *Aust. N. Z. J. Psychiatry*, vol. 54, no. 8, pp. 850–852, Aug. 2020, doi: 10.1177/0004867420943943.

[6] S. Lalitha and S. Tripathi, "Emotion detection using perceptual based speech features," in *2016 IEEE Annual India Conference (INDICON)*, Dec. 2016, pp. 1–5. doi: 10.1109/INDICON.2016.7839028.

[7] S. Basu, J. Chakraborty, A. Bag, and Md. Aftabuddin, "A Review on Emotion Recognition Using Speech," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Mar. 2017, pp. 109–114. doi: 10.1109/ICICCT.2017.7975169.

[8] U. Mahesh Yadav Konangi, V. R. Katreddy, S. K. Rasula, G. Marisa, and T. Thakur, "Emotion Recognition through Speech: A Review," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May 2022, pp. 1150–1153. doi: 10.1109/ICAAIC53929.2022.9792710.

[9] R. S. Sudhakar and M. C. Anil, "Analysis of Speech Features for Emotion Detection: A Review," in *2015 International Conference on Computing Communication Control and Automation*, Feb. 2015, pp. 661–664. doi: 10.1109/ICCUBEA.2015.135.

[10] N. Kamaruddin and A. Wahab, "Human behavior state profile mapping based on recalibrated speech affective space model," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 2021–2024. doi: 10.1109/EMBC.2012.6346354.

[11] A. Hanjalic, "Extracting moods from pictures and sounds: towards truly personalized TV," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 90–100, Mar. 2006, doi: 10.1109/MSP.2006.1621452.

[12] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, "Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference," *Circuits Syst. Signal Process.*, vol. 39, no. 9, pp. 4459–4481, Sep. 2020, doi: 10.1007/s00034-020-01377-y.

[13] J. Ye, X. Wen, Y. Wei, Y. Xu, K. Liu, and H. Shan, "Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition." arXiv, Aug. 14, 2023. Accessed: Aug. 18, 2023. [Online]. Available: http://arxiv.org/abs/2211.08233

[14] P. D. Paikrao, A. Mukherjee, D. K. Jain, P. Chatterjee, and W. Alnumay, "Smart Emotion Recognition Framework: A Secured IoVT Perspective," *IEEE Consum. Electron. Mag.*, vol. 12, no. 1, pp. 80–86, Jan. 2023, doi: 10.1109/MCE.2021.3062802.

[15] M. İleri and M. Turan, “Sentiment Analysis of Meeting Room,” in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Jun. 2021, pp. 1–5. doi: 10.1109/HORA52670.2021.9461354.

[16] T. Nguyen, R. Khadka, N. Phan, A. Yazidi, P. Halvorsen, and M. A. Riegler, “Combining datasets to increase the number of samples and improve model fitting.” arXiv, May 16, 2023. doi: 10.48550/arXiv.2210.05165.

[17] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[18] V. Singh and K. Sharma, “Empirical Analysis of Shallow and Deep Architecture Classifiers on Emotion Recognition from Speech,” in *2019 6th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/ 2019 5th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, Jun. 2019, pp. 69–73. doi: 10.1109/CSCloud/EdgeCom.2019.00-16.

[19] G. Luo, “A review of automatic selection methods for machine learning algorithms and hyper-parameter values,” *Netw. Model. Anal. Health Inform. Bioinforma.*, vol. 5, no. 1, p. 18, May 2016, doi: 10.1007/s13721-016-0125-6.

[20] D. Verma and D. Mukhopadhyay, “Age Driven Automatic Speech Emotion Recognition System,” in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Apr. 2016, pp. 1005–1010. doi: 10.1109/CCAA.2016.7813862.

[21] S. R. Kadiri and P. Alku, “Excitation Features of Speech for Speaker-Specific Emotion Detection,” *IEEE Access*, vol. 8, pp. 60382–60391, 2020, doi: 10.1109/ACCESS.2020.2982954.

[22] J. C. Vásquez-Correa, N. García, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, “Emotion recognition from speech under environmental noise conditions using wavelet decomposition,” in *2015 International Carnahan Conference on Security Technology (ICCST)*, Sep. 2015, pp. 247–252. doi: 10.1109/CCST.2015.7389690.

[23] R. Barber *et al.*, “A Multirobot System in an Assisted Home Environment to Support the Elderly in Their Daily Lives,” *Sensors*, vol. 22, no. 20, p. 7983, Oct. 2022, doi: 10.3390/s22207983.

[24] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, “Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study,” *IEEE Trans. Affect. Comput.*, pp. 1–1, 2022, doi: 10.1109/TAFFC.2022.3143803.