Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Multi-emotion Recognition and Dialogue Manager for VR-based Self-attachment Therapy

Author:
Lucia Simkanin

Supervisor:
Prof. Abbas Edalat

Submitted in partial fulfillment of the requirements for the MSc degree in Computing Science of Imperial College London

September 2020

Abstract

Despite increase in mental health awareness, there is still a large portion of people not receiving treatment. There are various side effects associated with drugs and not everyone has access to therapy, whether due to cost or availability. Self-administrable therapies allow for greater ease of access and as such, a novel virtual reality based self-attachment therapy is currently being developed. One of its main goals is for the user to develop a bond and reparent a child avatar representing one's childhood self. To do so efficiently, the emotions, felt by the user when discussing their childhood, should be transferred to this child avatar.

This project aims to train an emotion recognition model to allow the detection of multiple emotions. In addition, it aims to implement the self-attachment theory scenario to manage the dialogue between a user and a virtual therapist. The two parts of the project are to be integrated within the proposed virtual reality application.

The devised multimodal and multi-emotion recognition model trained on CMU-MOSEI dataset, achieves state-of-the-art results on happiness detection and competitive results for other emotions. The scenario was implemented using FAtiMA toolkit and together with the emotion recognition model was successfully integrated within the virtual reality application.

Acknowledgments

I would specifically like to thank my supervisor Prof. Abbas Edalat for their insight, advice and guidance regarding this project.

Additionally, I would like to thank Georgios Rizos for the insightful discussions, suggestions and feedback regarding the emotion recognition model.

I would also like to thank James Tavernor for providing me their model and framework to expand upon and Neophytos Polydorou for the smooth collaboration regarding the integration.

Overall, I would like to thank all who participated in our weekly meetings for their support.

Contents

Chapter 1

Introduction

This MSc project is part of larger group effort organised by Prof. Abbas Edalat to develop a virtual reality (VR) administered self-attachment therapy (SAT) application that should have the potential to treat certain mental disorders such as depression or anxiety. To understand how the individual project is integrated within, firstly the context and rationale of the group effort will be outlined, followed by an introduction of my part in the task.

1.1 Group Setting

The impact of untreated mental illnesses can be felt on multiple levels, especially due to their high prevalence (Demyttenaere et al., 2004), with approximately 36% to 50% cases with a serious condition left untreated in developed countries and 76% to 85% in undeveloped countries. As a result, individuals can experience various physical, social and emotional difficulties, negatively impacting their quality of life (Candilis & Pollack, 1997). Furthermore, the impact associated specifically with stigma and discrimination related to these illnesses, is observable also at the economic level (Sharac et al., 2010). It is clear that there are great costs associated with untreated mental health conditions, and therefore it is of importance to identify and address the reasons behind the lack of treatment.

There are several likely explanations as to why, individuals do not or can not seek professional help. To illustrate, on an individual level, the belief that professional help is only applicable to conditions associated with biomedical explanation, self-stigma and self-blame, often contributes to lowered willingness to seek help (Stolzenburg et al., 2019). On broader level, the misallocation of resources, such as preferential treatment of mild cases, may result in less resources available for more serious cases. Additionally, the issue lies within drugs being prescribed instead of therapy and the lack of professionally trained therapists (Laynard et al., 2007). There are further possible causes such as the high cost of therapy and its long duration in order to be effective (Honyashiki et al., 2014). Therefore, it is apparent that the main contributor is the lack of access to a professional therapy associated with cost, time, availability and stigma.

The impact of mental illness is even further prominent during the current situation associated with COVID-19. On one hand, individuals with mental disorders are unable or restricted to seek out help as to limit the spread of the virus. This exacerbates the already existing lack of access to conventional therapies, and if not addressed it may have a lasting impact on the society and economy in the future. On the other hand, the presence of the pandemic may increase the risk for mental illnesses, negative psychosocial outcomes and emotional distress even in the previously healthy population (Pfefferbaum & North, 2020). Therefore, there is a need to address these issues, especially as it is not known how long the pandemic may last, and to mitigate the impact of any similar situations in the future.

SAT is a recently developed self-administrable psychotherapy, which is heavily based on the notion of attachment theory (AT). AT has been formulated by John Bowlby and is based on the idea that human infants form an affectional bond, also called attachment with their main caregiver as an evolutionary mechanism to promote survival (Bowlby, 1969). SAT attempts to adjust one's attachment by steering an individual to love their childhood self, the same way as a parent would provide emotional support, nourishment and love to their child to stimulate a secure attachment (Edalat, 2015). It does so in the following four steps; Firstly, an individual is introduced to the scientific basis and hypothesis of the therapy, to become motivated and committed to practising the therapy. Secondly, one attempts to connect compassionately with their childhood self, by visualising themselves as a child and interacting with the child. Thirdly, falling in love with the childhood self is achieved by imagining a ceremony where they promise to take care of the child as a parent would, encompassing singing a song to trigger happy memories of their childhood. Lastly, re-parenting the childhood self, to facilitate a secure attachment by minimising any negative emotions and maximising positive ones.

Utilising VR as the environment for a self-administered therapy has several advantages. One does not need to search for an available therapist and travel to their location but can now receive treatment from home. Moreover, VR equipment is only associated with one-time payment which is likely to be lower than paying for several therapy sessions, therefore reducing the cost associated with therapy. Home administered therapy can also potentially reduce the stigma associated with seeking professional help. Overall, it can be seen, that self-administered VR therapy allows for ease of access, which is the main challenge in seeking professional help.

Due to the reasons portrayed above, SAT is a reasonable candidate to be administered in VR and is currently being developed (Cittern et al., 2017) and will be in the soon future tested in clinical setting (Ghaznavi et al., 2019). Another SAT-specific advantage when considering VR is the immersion achieved. In the normal procedure one needs to utilise their imagination to some extent to progress through the stages. Despite the visual aid of childhood pictures when envisioning themselves as a child, the participant still needs to employ a level of imagination to interact with

the child. However, there are individual differences in mental imagery (Phillips, 2014), whether one considers the ability or the enjoyment involved. As such, some people may struggle with the tasks explored in SAT, but not in VR administered SAT, as there is limited need for mental imagery.

As mentioned above, one version of VR SAT is currently in development, this version is primarily to be used with a mobile phone and a cardboard VR headset, which is the most affordable option. However, due to the virtual environment being emulated by a mobile phone, this application lacks interactivity and the graphics are less realistic, than when using a higher end VR headset with controllers. Therefore, a more interactive version is to be developed within this group project, utilising a higher end VR headset, more specifically the Oculus Quest. The advantages of this headset include its lower price when compared to other high-end headsets, and at the same time it is capable of emulating realistic virtual environments. Moreover, it is untethered which means that the headset does not need rely on computer for processing power and therefore there is greater degree of freedom when moving around in the VR environment. As well as, its hand tracking features which allow for the user to interact with the environment using hands rather than controllers, resulting in a more innate interaction.

1.2 Individual Contribution

The aim of this project is to develop an emotion recognition model that should be able to detect several human emotions. Furthermore, this model should be integrated within a larger framework that will be used to implement the interaction in the form of dialogue between a user and a virtual therapist, whose purpose will be to guide the user through SAT.

The rationale behind the emotion recognition model is the following, to develop a highly interactive and user-tailored VR SAT experience, we will need to recognise the user's current emotional state. More specifically, the application will consist of several protocols in which the user is to interact with the childhood self and these often involve projecting the user's emotions onto the child avatar. By asking the user about their feelings regarding certain situations and then having the child avatar visibly feel a similar emotion, the user can better relate to the situation. Therefore, they should be able to learn how to comfort and reparent their childhood self more efficiently, as well as immerse in the therapy and the environment to a greater degree. Additionally, tracking the user's emotions at certain points in the therapy will allow to compare the progress they have made, which could be used to suggest certain protocols the user should try.

Considering the rationale for including the virtual therapist, the major limitations of the mobile-based VR application, as identified in a usability evaluation of the platform (Ghaznavi et al., 2019), included the lack of navigation through the therapy, as well as the limited explanation of the various protocols. To take these into account when developing the new application, a decision was made to include a therapist avatar that would guide the user through the therapy. Additionally, the player would also engage in a scripted dialogue with this virtual therapist to describe their feeling regarding certain situations. Throughout the interaction, the virtual therapist would try to evaluate the user's responses, using the emotion recognition model to detect any underlying emotions felt by the user. Furthermore, due to the inclusion of the virtual therapist, we will be able to ask for confirmation of the results produced by the emotion recognition model to avoid any incorrect detection. Overall, the inclusion of the virtual therapist should increase the interactivity and immersive nature of the application.

1.3 Outline

In this report, the process of achieving the presented aims will be described in the following chapters. Firstly, a more detailed account of the concepts mentioned will be given in Background, as well as a description of appropriate techniques and related works. Secondly, the design choices and the technical implementation will be detailed in the chapters Emotion Recognition Model and Virtual Therapist Integration. Thirdly, the achieved results will be discussed in Evaluation, regarding both the emotion recognition model and the virtual therapist integration. And finally, the report will be concluded in the last chapter, accompanied with a discussion of future improvements.

Chapter 2

Background

2.1 Mental Health

2.1.1 Mental Disorders

A mental disorder, often referred to as mental illness, is a hard concept to define, as there does not exist a formal definition that applies across all contexts. A common factor across most of the efforts to define a mental disorder, is that the mental or behavioural condition causes dysfunction (Stein, 2013). To further specify this, one should turn to official classification manuals, that attempt to term mental disorders and categorise them. It is worth noting, that the content of these manuals is subject to evaluation and change and therefore the concept of mental disorder is somewhat fluid and time dependent. Currently, there are two reliable manuals that professionals turn to, to guide them when assessing whether one suffers from a mental disorder, the 5th edition of The Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2013), also known as DSM-V and the 10th revision of The International Statistical classification of Diseases and Relate Health Problems (World Health Organization, 1992), referred to as ICD-10.

DSM-V is published by American Psychiatrist Association and is the preferred manual used by clinicians when diagnosing mental disorders in the United States. DSM-V defines a mental disorder based on several features (Stein et al., 2010). To illustrate, the condition should be of behavioural or psychological origin and cause significant distress or impairment and is not the outcome of the individual's reaction to the stressful environment, such as a death of a relative. Furthermore, the DSM-V considers that no single definition can encompass the variety and variance of mental disorders and therefore the manual should act only as a guide and the final judgement rests on the clinician. DSM-V groups known mental conditions into categories either based on the underlying factors, such as Neurodevelopmental disorders, or based on common symptoms such as Depressive or Anxiety disorders. Additionally, instead of a strict definition per disorder, it is characterised by number of features that usually occur in such disorder. Often, it is not a binary decision of whether someone is ill, but rather to what extent they fall on the spectrum of symptoms.

ICD-10 is devised by the World Health Organisation and its use is more prevalent internationally. In comparison to DSM-V, it includes not only mental disorders but also other non-mental diagnoses. ICD-10 mental disorder definition focuses mostly on distress and personal life impairment related to the clinical symptoms (IAG for the revision, 2011). In comparison to DSM-V the ICD-10 favours the simplicity of a mental disorder definition. ICD-10 groups disorders based on multiple factors, such as onset and the underlaying basis of symptoms or a disorder, whether neurological or behavioural. Additionally, each disorder-specific definition consists of main clinical features and other less specific features. Similarly to DSM-V, the diagnostic guidelines allow for the clinician to make the final judgement, rather than purely rely on the manual.

The two very prevalent mental disorders which could benefit from SAT, are anxiety and depression disorders. Anxiety disorders encompass a range of diagnoses and the most characterising symptoms include feelings such as worrying, rumination and fear about present and future occurrences (American Psychiatric Association, 2013). The disorders vary based on what the feelings are directed towards and include generalised anxiety disorder, phobias, social anxiety etc. Depressive disorders are characterised based on usually recurrent episodes of depressive feelings, loss of pleasure from activities, sometimes irritability and their effect on individual's functioning (American Psychiatric Association, 2013). Based on the specific symptoms the disorders vary, however the most common is major depressive disorder. It is also important to mention, that often anxiety and depressive disorders can co-occur quite frequently (Gorman, 1996) causing further distress to the individual. In terms of treatment, there are various options, involving medication and therapies, with therapies such as cognitive behavioural therapy being often recommended. This is especially true for the less severe cases, due to the lack of side effects, as associated with medication and the therapy's efficacy at treating both anxiety disorders (Norton & Price, 2007) and depression (Gartlehner et al., 2017). However, as mentioned in the introduction, there are various disadvantages associated with conventional psychotherapies, especially the cost and lack of availability, which could be addressed by self-administrable therapies.

2.1.2 Self-attachment Therapy

Attachment is an affectional bond between a child and their caretaker that potentially offers the evolutional benefit of survival for the child (Bowlby, 1969). A child therefore seeks to form an attachment with any parental figure through interaction with them during their early childhood, especially those who are sensitive to their needs. Once an attachment is formed, the child begins to think of the parental figure as their safe place, and once a stressful environment is encountered the child will look for their caregiver. Furthermore, on the basis of the attachment, an infant creates an internal working model, which will serve as a groundwork, for the future adult interactions with the world (Bretherton & Munholland, 2008). The internal model is formed based on how the caregiver reacts to various situations with regards

to the child and the stressful environment.

Based on a Strange Situation Procedure designed by Mary Ainsworth, which is a way of assessing infant's reaction to stress and the mitigation including their caregiver, it was observed that infants form different attachment styles with their primary caregiver (Ainsworth et al., 1978). The procedure consists of various situations which allow for observing the infant's reaction to them. It begins with the child and their caregiver entering a room. Firstly, child is allowed to explore the room while the parent does not interfere but is in the room with the child. Secondly, a stranger comes into the room and starts interacting with the caregiver first and then the child. Thirdly, parent leaves the room and after a brief period of time re-enters and interacts with their child, while the stranger leaves. Fourthly, the parent leaves again and the child is alone in the room, which is followed by the stranger coming back and again interacting with the child. Lastly, the parent comes back, comforts the child and the stranger leaves.

During the procedure, various behaviours of the child are observed, such as how explorative the child is, their reaction to the stranger when with parent and when alone, as well as their reaction when the caregiver leaves and comes back and comforts the child. Based on the above, the child is classified as having one of the four attachment types which include secure, anxious-ambivalent, anxious-avoidant and disorganised attachment.

A child is securely attached when they are happy to explore the room and interact with the stranger when the caregiver is present. Additionally, they are distressed when the caregiver leaves but are also easily comforted when they come back. Anxious-ambivalent attachment is an insecure form of attachment, which results in the child feeling distressed even when the caregiver is in the room and the caregiver cannot comfort the child easily. Upon their return, the child either shows feelings of resentment or helplessness. Another insecure type is anxious-avoidant attachment, where the child does not like to explore the room and is either ignoring the caregiver or tries to avoid them. Moreover, they do not seem to care whether the caregiver leaves or comes back to the room. Finally, the disorganised attachment was discovered later (Main & Solomon, 1990) and refers to a child which does not display either ambivalence or avoidance, however they can show fear, freezing or uncoordinated movement as a reaction to the caregiver departing. Furthermore, it is more of an open-ended category and often children who showed a mixture of the insecure attachment behaviours were classified as disorganised. The type of attachment can influence one's adult life, more specifically insecure attachments can increase the risk for developing mental disorders, such as depression (Murphy & Bates, 1997) or anxiety (Eng et al., 2001). Additionally, different attachment styles affect one's emotion regulation, such as their response towards stressful situations (Mikulincer & Shaver, 2019).

As mentioned before, an insecure attachment can have an adverse effect on one's adult life. SAT (Edalat, 2015) introduces the notion that even if one experienced an insecure attachment as a child, they may be able to reform the attachment to secure. Moreover, if the root or risk factor for adverse events in adult life is insecure attachment, this reformation should also reduce or mitigate the impact of the negative implications. Therefore, SAT therapy can treat some mental disorders such as depression, which have heightened risk of developing in an individual with insecure attachment as a child. To tackle the change in attachment, the individual through mental representation takes the place of the caregiver to their childhood self. To assist in imagining the child, the individual uses their childhood pictures. Following, various protocols are experienced in order to emulate the secure attachment between the individual and their childhood self, hence the name self-attachment.

To assess whether SAT has the potential to treat mental health conditions such as depression or anxiety, there is neurobiological, psychological and computational indirect evidence to support its use as psychotherapy, as well as success in pre-clinical trials (Edalat, 2017). However, it should be noted, that SAT is still in the early stages in comparison to already established psychotherapies, which should be considered when assessing its efficacy. The potential and theory to support SAT is present, and the preliminary results are in favour of its use as psychotherapy, indicating that it is worth researching further. Although, other therapies have the advantage of being researched extensively, perhaps the main advantage of SAT in comparison to traditional therapies is that it is self-administrable, greatly improving the ease of access to most people. In addition, the pilot studies show that SAT had a positive impact even on cases which did not respond to the traditional therapies.

2.1.3 VR-administered Therapy

Due to the advancements in technology, VR can emulate real world setting well without the need of a high-end performing computer. As such, even with the help of a smartphone one can interact with virtual environment in real time quite persuasively, with the immersiveness only increasing with the use of more powerful computing devices. Due to the prevalent use of mobile phones and the relatively low cost associated with the VR equipment, which is likely to get lower in the future, VR seems to be an appropriate platform to explore in terms of mental health treatment. Furthermore, it has potential to present multimodal stimuli in an environment mimicking the real world, which allows for more controlled experiments and safer experience, while still maintaining high levels of ecological validity.

Up to this date, VR has been employed in a variety of settings including therapy and rehabilitation. In terms of its health-related potential and efficacy, it has been identified as desirable in assessing and treating brain injuries (Davies et al., 1998) and in rehabilitation of neurological disorders (Rizzo et al., 1998). Furthermore, when considering mental disorders, it has proven to be at least as effective in treating various phobias through exposure, and even more effective dependent on the phobia

(Emmelkamp et al., 2001). Moreover, VR is compatible with neuroimaging techniques which further increases its potential in assessing or treating neurobiological disorders (Romano, 2005). Moreover, there is evidence that VR administered therapy is effective at treating depression (Romano, 2005), anxiety and post-traumatic stress disorder (Difede & Hoffman, 2002). However, it should be taken into consideration that most studies exploring the efficacy of VR therapies do not utilise the best research practises (Fodor et al., 2018), whether in terms of conducting the experiment or reporting, and further independent and higher quality research should be conducted.

2.2 Emotion Recognition

2.2.1 Emotions

With the task in mind, one has to specify what it means to recognise emotions, especially what constitutes an emotion and which emotions should be detected. Despite the amount of emotion research conducted, there is not necessarily a converging opinion on what an emotion is. There are various theories that encompass the notion of emotion and these can be grouped into three categories, based on the possible causes of emotions (Moors, 2009). Firstly, an emotion is caused by a body's physiological response to environmental stimuli. Secondly, an emotion is caused on neurological basis. And lastly, cognitive processes such as thoughts give rise to an emotion.

Perhaps, rather than trying to define the concept of emotion, within this project it is more important to explore emotion classification. One widely accepted theory regarding emotions is that six basic emotions emerged due to evolution: happiness. surprise, fear, sadness, anger, and disgust (Ekman, 1992). The term basic refers to the universality of such emotions, meaning that they can be expressed facially and understood even across cultures. Other categorical models can include a subset of the six basic emotions or can consist of some of the later identified universal emotions, such as pride or shame (Cordaro et al., 2018).

In contrast to the categorical model, various dimensional models to identify emotions have been proposed. A dimensional model usually expresses an emotion as a point on various axes. One of the widely accepted 2D models is the circumplex model which is composed of a circular 2D plane, with the horizontal axis representing valence and the vertical axis representing arousal (Russell, 1980). Furthermore, one could also consider 3D models, such as the PAD emotional state model which classifies an emotion using three axes: pleasure, arousal and dominance (Mehrabian & Russell, 1974).

As can be seen, there are various models of emotion classification and prior to any emotion recognition, one must decide which models should be considered. For the purpose of this project, the categorical model consisting of six basic emotions was selected. The choice was mainly affected by the following factors. Firstly, as the

aim is to implement an emotion recognition model which can be used in the SAT VR application, fear is a desirable emotion to predict as it can be a symptom of anxiety, which SAT can potentially treat. Secondly, the choice was affected by the dataset selected, as described below, which contains only these six basic emotions.

2.2.2 Modality

Another design choice to be made when considering emotion recognition is the modality of the data. More specifically, it depends on whether one is interested in predicting emotions from audio, text or visual data. Another possibility is to approach the problem multimodally, by utilising a combination of the various data. With regard to this project, the following considerations were made. As the device responsible for acquiring the data from the user of the SAT VR application is Oculus Quest, its capabilities need to be considered. The device is capable of recording audio and this audio can be transcribed into text independently of the device. However, the device is not able to record a video of the user, as well as no external camera can be used as the headset is covering the user's face. Therefore, the modalities that can be used within this project are limited to audio and text. However, it is worth mentioning that recording video of a user's face through a VR headset will be a possibility in the future (Heaney, 2019). Based on whether such feature will be accessible on a commercial headset, the visual modality could be considered for future implementation.

2.2.3 Machine Learning

Machine learning (ML) is a subfield of artificial intelligence (AI), concerned with automatic improvement of computers through experience. It uses some form of training data to learn about certain relationships in the dataset. Its goal is to train a model that can make predictions, with the model performing well on new unseen data. The most common ML types include unsupervised, supervised and reinforcement learning. For the relevance of this project supervised learning will be explored. To define, supervised learning incorporates both input and output data in its training phase. More specifically, it tries to learn the relationship between certain input, in our scenario some text and the associated output, emotion. After the training phase, the learnt model is given only input data without the actual output and is tasked to predict the output. However, to make reliable predictions a large amount of training data is required for the model to learn, which is one of the biggest drawbacks of ML.

2.2.4 Natural Language Processing

Natural language processing (NLP) incorporates the fields of computer science, AI and linguistics to allow computers to assess and utilise data related to human languages. In terms of this project, the major challenge will be natural language understanding, specifically emotion recognition. Furthermore, NLP can be separated into rule-based NLP, which involves manually crafted rules, and statistical NLP, which

utilises ML. Statistical NLP is the preferred method to employ as of today due to its automaticity, efficiency and scalability. NLP utilises various syntax related tasks, which can be of use on their own and/or can aid each other in the bigger picture and are useful for pre-processing the data. These tasks include lemmatisation, stop words removal, tokenisation, parsing, part of speech tagging, stemming etc.

2.2.5 ML Architectures

Artificial Neural Network

An artificial neural network (ANN) is a collection of artificial neurons, which are units connected together in various layers, based on the biological neurons in brain. Each neuron can receive, process and pass along a signal. In the training phase, the input neurons receive a signal, which they process and propagate forward through all of the other neuron layers (hidden layers) until the output layer is reached. Following, the output from the network is compared to the actual output ,and through a scheme called backpropagation, the parameters associated with the neurons are adapted so that the predicted output matches the actual. In the test phase, the network receives the data without the actual output, propagates only forward and predicts the output. Based on how the predicted output and actual output differs on unseen data, one can evaluate the performance and generalisability of the model. To achieve a better performance the training phase is usually done in batches, where only a certain set of the training data is given to the network at a time and after which the network updates its parameters. Additionally, it is worth to note that for complex problems a network might require multiple layers to optimally learn, which is a process called deep learning.

Recurrent Neural Network

An often-employed type of neural network in NLP and processing audio data is recurrent neural network (RNN). RNN is similar to ANN, with the difference that RNN includes an internal memory. Consequently, RNN is able to process sequential data such as sentences, whereas ANN cannot remember any previous input experience. Due to the importance of context in natural language understanding, RNN is specifically suited for emotion recognition tasks. However, a certain disadvantage is that RNN has problems with modelling the relationship between two inputs in a sequence very far apart.

Long Short-term Memory Model

To address the shortcomings of RNN, long short-term memory (LSTM) model has been invented. As a variant of RNN, LSTM inherited the ability to remember and have enhanced it to be able to remember even longer time periods. In addition, they possess the ability to forget information, therefore they can selectively remember and forget any useful or not useful information. This is especially an important feature,

as emotions may be context dependent and therefore an architecture is needed that is able to capture the context.

Transformer

The transformer is a recently introduced deep neural network model (Vaswani et al., 2017), not too different to RNN in function. The similarity stems from the ability to process sequential data, which is specifically useful in NLP. However, there are some limitations to RNN, that a transformer can address. To illustrate, RNN can only process sequential data either left-to-right or right-to-left, and therefore may be inefficient if some data are further away from each other in a sequence. Whereas, a transformer takes into account the relationships of all of the datapoints in a sequence without being limited to their positions. Furthermore, transformers are able to process data in parallel resulting in shorter training time, in comparison to RNN which can only process data sequentially. Therefore, transformers are computationally more effective and are currently considered the state of the art in NLP.

Bidirectional Encoder Representation from Transformers

Currently, the most employed model in NLP is the Bidirectional Encoder Representation from Transformers (BERT). Published recently, its main intention is "... to pretrain deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers" (Devlin et al., 2018, p. 1). Therefore, one is able to finetune a previously trained BERT model to suit their task. At the time of the publishing, BERT achieved unprecedent performance on various NLP tasks. Due to the above, it would be beneficial to consider utilising BERT in emotion recognition.

BERT in comparison to other models does not need significant amount of text preprocessing as it only requires word-level tokenisation and encoding of the data to process it. When considering the task of emotion classification, tokenisation consists of dividing the sentence into word tokens and adding a [CLS] token to the start of the sentence which is accessed to depict the sentence itself. Furthermore, a [SEP] is added to the end of the sentence to mark its end. Additionally, to process the sentences with BERT, they need to be of the same length, and therefore for any sentence that has less words than the selected length [PAD] tokens need to be added to make up for it. As a previously trained BERT is used, the tokenised sentence needs to be transformed into unique id's as that is what the original BERT is trained on. The unique id's are predetermined based on the vocabulary that BERT was trained on and if an unknown word is encountered it is given the token [UNK].

2.2.6 Databases

As ML requires a large amount of data to reliably train a model, a well-suited dataset needs to be selected. The choice of dataset will also closely impact the emotion classification model. Therefore, various datasets suitable for emotion recognition

are discussed below, as well as the final decision of which dataset is to be used and the reasoning. All of the datasets mentioned contain both the speech and text modality at minimum.

IEMOCAP

Busso et al. (2008) have constructed the Interactive Emotional Dyadic Motion Capture database, which is an audio-visual database intended for emotion recognition. In addition to the audio-visual aspect it contains transcribed text data, which is segmented into sentences. The content makes up for approximately 12 hours, consisting of approximately 7532 utterances. To collect the data, 10 actors engaged in 5 conversation sessions, including a second party, which were either scripted or improvised to display various emotions. In addition to the visual and audio recordings, the actors wore face markers to track their face movement. In terms of emotional annotation, two methodologies are employed. Labels are given based on categorical approach to emotions: neutral state, happiness, sadness, anger, surprise, disgust, frustration, excitement and other. However, it should be noted that disgust, surprise and fear are largely underrepresented. In addition to the above, valence, arousal and dominance scores are assigned to the data.

RAVDESS

Livingstone and Russo (2018) have established the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This database contains both the audio and visual modality, with the text data consisting of a set of sentences. All recordings consist of one utterance, with 7356 utterances in total. To collect the data, each of the 24 professional North-American actors performed 104 unique scripted recordings. The emotions they were asked to display are: happiness, sadness, anger, fear, surprise, disgust, calm and neutral. In addition, each of the emotions was displayed using normal and string intensity. The dataset was also validated by 247 independent research participants, to make sure that the actors' performance was genuine and that the emotions portrayed consisted of the desired emotion and intensity.

An interesting feature of this dataset is that, in addition to the speech data outlined above it contains emotional singing as well. The singing part of the dataset consist of the following performed emotions: happiness, sadness, anger, fear and calm. This feature could be of interest to this project as a certain SAT protocol involves singing to the child, and therefore if this dataset was used, emotion recognition could be performed at that time as well.

CREMA-D

Cao et al. (2014) have created the Crowd-sources Emotional Multimodal Actors Dataset (CREMA-D), which is of similar nature to the RAVDESS dataset, as it also contains an actor performed emotional data, specifically audio and video. Overall,

7442 utterances were produced portraying the following emotions: happiness, sadness, anger, fear, disgust and neutral. For this dataset, 91 actors were employed of various genders, ages and ethnic backgrounds. To collect the data, scripted sessions consisted of 12 sentences acted out to display the above-mentioned emotions. In addition, 2443 people rated the emotions and their intensity to validate the dataset and obtain intensity ratings.

CMU-MOSEI

Zadeh et al. (2018) provide the largest multimodal dataset for emotion recognition, including the forms of video, audio and text, the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). The database contains a total of 23453 sentences, annotated in terms of the six basic emotions with happiness and fear being the most and least prevalent classes. In comparison to the other mentioned datasets, CMU-MOSEI is not a product of acted out sessions but rather a collection of 3228 real-life YouTube videos, consisting of 1000 distinct speakers discussing approximately 250 topics forming over 65 hours of content. To obtain the text data, the videos chosen for this dataset must have included properly punctuated manual transcriptions. The annotation was performed by 3 research participants and instead of classifying a utterance as one of the emotions, each utterance was rated for each emotion on a 3-point Likert scale with the following meanings: 0 – no emotion, 1 – weak presence, 2 – moderate presence, 3 – high presence of an emotion.

This database was chosen as the most appropriate for our task, for these reasons. Firstly, the major advantage of CMU-MOSEI in comparison to the other databases is that it is not a product of acting. Therefore, as it consists of real-life videos, we can expect that with the use of this database the emotion recognition model should have higher external validity. This is specifically important, as the intended purpose of the emotion recognition model is to be used in the SAT VR application. Consequentially, as the input from this application to the model will be real-life data, training the model on other real-life data should allow for better generalisation. Secondly, there is a certain disadvantage for both the RAVDESS and CREMA-D dataset. The scripted sentences that are read by the actors are completely neutral in terms of any emotional value e.g. It's eleven o'clock (Cao et al., 2014). As we are interested in both audio and text modality, these two datasets are less desirable. Thirdly, CMU-MOSEI is quite large with even the least prevalent emotion fear accounting for approximately 2000 utterances, which is an important feature as larger dataset allow for models with better performance and validity.

2.3 Virtual Therapist Framework

With regards to the second part of the project, regarding the virtual therapist, a good framework should be chosen that will allow for the implementation of the dialogue between the user and the therapist. A desirable feature of such framework should be the ease of use. This is important not necessarily for the scope of this project

but mainly for any future extensions. More specifically, we can expect that the SAT protocols could evolve in the future, with possible addition of new protocols or even a change of wording in the current ones. The selection process is described below, including the possible options and the reasoning behind selecting one.

2.3.1 Virtual Human Toolkit

Virtual human toolkit (VHT) was developed by Gratch et al. (2013) and was inspired by cognitive psychological research to provide an easy tool for researchers to create virtual humans (VH) and to expand the research in this area. The toolkit consists of various modules in the area of NLP, speech recognition, non-verbal behaviour generation and etc. It relies on Unity as its engine to emulate VH in virtual environment. Rather than a framework, VHT is an application that allows the user to select a small variety of environment such as college campus or a house and to include VH in such scenes. Additionally, in terms of dialogue management, phrases, questions and answers are paired together using a classifier, which results in the VH responding to the player, using the pairings. Moreover, non-verbal behaviour animations can be added, whenever the VH says a certain word.

Overall, these features are very promising and perhaps if one was developing an application alone, this toolkit could be desirable to use. However, for the purpose of this project, which is to be incorporated within a larger group project, there are several disadvantages regarding VHT. Firstly, there is no module that would allow emotion recognition model to be implemented. Although, the authors mention that VHT can be used as a library or that one can develop their own module, there is a lack of documentation that would help in achieving this. Therefore, it is quite hard to navigate through their code to understand how an emotion recognition model could be integrated within the toolkit. Secondly, as mentioned the scenes are already implemented using Unity, however the toolkit can only be run through an already compiled application. This is undesirable as within the group project, the virtual environment and avatars are already being implemented. One could consider merging these two Unity projects, but this is not possible as the authors have stopped providing the Unity project used within the toolkit. Thirdly, VHT does not allow implementing a guided scenario consisting of several steps that are consequentially played out. This is especially important, as the SAT protocols and the overall scenario are already established, while using VHT would only allow for pairing questions and answers together.

2.3.2 FAtiMA Toolkit

FAtiMA toolkit was created by Mascarenhas et al. (2018) and it is an open-source project for developing virtual agents for serious games. Not only it is a capable dialogue management engine, but it has other capabilities such as scenario creation, emotional decision making, autobiographical memory and etc. The toolkit can be used in various ways, for example, a C# library, only exploiting the features nec-

essary for a project. Additionally, the Integrated Authoring tool provides a simple visual interface, through which the various scenarios, rules and variables can be created. One can also benefit from using the tool to create a scenario which is then implemented along with any other modules necessary. As FAtiMA is implemented in C#, it allows for an easy integration with the Unity engine. However, a disadvantage to certain projects is that FAtiMA does not include any avatars as well as no text-to-speech capabilities that would result in a more interactive experience. As important as this may be for some developers, for the purpose of this project the virtual environment including the avatars and any of their behaviour is developed by another member of the group (Polydorou, 2020). Therefore, the framework chosen does not need to contain these features and should rather allow for easy integration within the VR platform.

After a careful consideration and experimenting with both toolkits, FAtiMA was chosen as the framework to be used when implementing the SAT scenario and integrating it with the emotion recognition module. The reasoning is as follows: firstly as mentioned, FAtiMA can without too much effort be used as a library which allows the integration of the emotion recognition model as well as the integration of the created scenario and the final virtual environment for the whole application. Secondly, the authors provide various well-explained tutorials and a more thorough documentation of their code. Consequently, one can expect that it will take considerably less time to familiarise oneself with the framework and allow for more time to be spent on other aspects of the project. Lastly, not only does the toolkit provide an easy way of incorporating scripted scenarios, it has additional capabilities, such as autobiographic memory which may be useful if not for this project then for any future implementation.

2.4 Related Work

2.4.1 Emotion Recognition from Text

Emotion Classification with Natural Language Processing (Comparing BERT and Bi-Directional LSTM models for use with Twitter conversations)

Joselson and Hallén (2019) have compared the use of BERT and their own novel architecture, concatenated word-emoji bidirectional LSTM (CWE-LSTM), for emotion recognition. They have participated in an NLP challenge SemEval, and their specific task was to create a model, that can classify the emotion of the last part of a twitter conversation. The emotions to be categorised were happy, sad, angry and other. The twitter conversations lasted three turns and each turn could have included either a sentence, emoji or the combination of both. CWE-LSTM model's input was made of word and emoji embeddings, each of which was processed by a separate LSTM layer, however for the word embeddings the layer was bidirectional. In total, 30160 conversations were used to train the model.

During evaluation, the CWE-LSTM model performed better in all 3 emotion categories then the BERT model, the difference was especially visible in the happy and angry category. Furthermore, the CWE-LSTM was trained under 90 minutes, which is significantly faster than the training time for BERT, which was slightly over a day. However, as the authors note, the difference was likely due to training on CPUs, and could be smaller if GPUs were utilised, as BERT can be parallelised on GPUs. The shortcomings of BERT in this paper were possibly related to overfitting, as well as due to the complexity of BERT architecture it was harder to understand why the model overfit the training data. Additionally, BERT struggled with learning from emoji and conversations consisting of shorter turns. On the other hand, the CWE-LSTM placed too much value on emojis, as well as swear words, sometimes failing to grasp the context of the situation. The authors advise the use of more direct approaches such as LSTMs with domain-specific data, rather than BERT.

Semantic-Emotion Neural Network for Emotion Recognition From Text

Batbaatar et al. (2019) proposed a novel architecture Semantic-Emotion Neural Network (SENN) to address the shortcomings of models not being able to detect the emotional relationships between words. SENN consists of two networks, bidirectional LSTM (BiLSTM) and convolutional neural network (CNN). The BiLSTM module is responsible for semantic encoding between words, by capturing both information flows, forward and backward. The CNN module has a function of emotion encoder, using convolving filters. To assess the performance of SENN, an emotion recognition task was introduced. In terms of the dataset, a part of a unified corpora (Klinger, 2018) was used, and the model was responsible for predicting one of the six basic emotions.

When evaluating the model, three different variations associated with different methods of word embeddings were compared. SENN performed better on almost all of the datasets used, outperforming other baseline models such as BiLSTM, LSTM and CNN+LSTM. In terms of execution time, SENN was slower than non-deep learning models, however in comparison to other deep learning models SENN was faster than BiLSTM and LSTM. The authors conclude that SENN performs better on average than the state-of-art models; but this conclusion could be challenged as a comparison with BERT is missing.

Practical Text Classification With Large Pre-Trained Language Models

Kant et al. (2018) perform an evaluation of pre-trained models such as transformers on an emotion classification task. Initially, a model is trained on unlabelled text data and then is transferred to solving supervised tasks. The task involved, is one the SemEval challenges with provided dataset of 6857 tweets, categorised into eight emotion categories: joy, sadness, anger, fear, trust, disgust, surprise and anticipation. Additionally, approximately 15000 tweets were collected relevant to a particular company to compare the performance on a domain-specific task. The initial

model was trained on a dataset consisting of Amazon reviews to include rich emotional context. After the pretraining is done, the model is finetuned on a specific task.

In terms of performance, the model outperforms a baseline comparison IBM Watson API, in the SemEval emotion recognition challenge. Additionally, it outperforms other previous submissions in the challenge, which were using BiLSTMs and training on much larger dataset. Considering the domain-specific task, the authors' transformer performs better than multiplicative LSTM as well as the Watson API. The authors note that a general-purpose API struggles with domain-specific tasks, as it could be impossible to produce context-independent emotion classification. The advantage of a transformer is that it can include greater variety of features relevant to different contexts and that with finetuning one can focus on the most important features. Overall, a transformer seems to be a flexible and effective framework in general, as well as specifically in emotion recognition.

ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT

Huang et al. (2019) describe a novel hierarchical LSTM for Contextual Emotion Detection (HRLCE) and compare its performance to BERT. The model was devised for the SemEval challenge, specifically the task concerning contextual emotion detection in text. As described before, 30160 twitter conversations of three turns were utilised to train the model to predict the emotion associated with the last turn in conversation. Specifically, the emotions to be classified are happy, sad, angry or other. Three types of word and emoji embeddings are inputted to the HRLCE. The hierarchical part of HRCLE is based on a hierarchical or context recurrent encoder-decoder (HRED) and was found to capture context well in dialogues. The HRED consists of two RNN units, an encoder RNN and a context RNN, which specifically allows to model the exchange in dialogue to capture context.

In terms of evaluation, the HRLCE model outperforms BERT, but only marginally. The difference in performance can be seen in the happy category, whereas in angry and sad category the performance does not differ on average. As this difference in performance is so small, it cannot reliably be concluded that the HRLCE outperforms BERT as the authors state. In addition, there is no mention of training time, execution time or any other variable that could help in distinguishing which one of the models is more advantageous to use. Nonetheless, for the purpose of the challenge the authors have combined both the HRLCE and BERT and ranked quite high in the competition.

EmoDet2: Emotion Detection in English Textual Dialogue using BERT and BiLSTM Models

Al-Omari et al. (2020) have built a novel architecture EmoDet2 for emotion detection. Specifically, EmoDet2 contains five sub-models: EmoDense, EmoDet-BiLSTM-submodel1, EmoDet-BiLSTM-submodel2, EmoDet-BERT-BiLSTM cased and uncased.

To illustrate, EmoDense is a classic ANN with four hidden layers, EmoDet-BiLSTM sub models utilise BiLSTM with different word embedding inputs and EmoDet-BERT-BiLSTM submodules have the raw data as input and rely on BERT to extract embeddings which are then inputted into BiLSTM layers. To authors have assessed their model using the SemEval contextual emotion detection in text task, as described before.

After training, the combination of four modules achieves a high accuracy for the given task, which is significantly higher performance than the baseline model provided by the SemEval organisers. However, no comparison is made to simpler models or any other complex models, whether considering performance or time efficiency. Therefore, it is impossible to conclude whether this model is efficient at detecting emotions, as it could be perhaps over-complicated or too time consuming.

Summary

Overall, when considering the above portrayed research, it is evident that the textual modality even on its own plays an important role in emotion recognition. Therefore, it would be worth considering, when choosing the modalities for this project. In terms of recognising emotions from text various model architectures are used, more specifically often Bi-LSTM and BERT are employed. In certain scenarios one tended to outperform the other, while in others this could not be concluded. Therefore, when choosing which one to use for processing the text data we should consider other factors as well. For example, BERT does not require too tedious data pre-processing but the raw textual input with only tokenisation and embedding. As extracting word-embeddings from the text requires a deeper level of understanding of NLP, it could be more time efficient to consider employing BERT rather than Bi-LSTM's. Although, it is also worth noting that BERT is more time consuming in terms of training time, but this could be improved upon by utilising training on GPUs and parallelisation.

2.4.2 Emotion Recognition from Audio

ADIEU FEATURES? END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NETWORK

Trigeorgis et al. (2016) introduced a novel network for emotion recognition using speech. Previously, most of the work on this topic mostly utilized a CNN and a LSTM models to process speech, together with extracted features from the speech as an input to these models. However, the authors set out to construct a network which could process the raw speech data without any pre-processing. Their model consists of convolutional layers which stand in for the role of the usually used features, followed by recurrent LSTM layers, with the model predicting arousal and valence. As raw data is fed as the input to the model, this approach is following the end-to-end principle.

To consider the efficacy of such model, RECOLA database (Ringeval et al., 2013) was used to train and test the model. The database consists of audio data collected using 46 French speakers who provided 5-minute recordings each, with additional modalities not explored in this paper. The results show that the new model performs significantly better than other models, trained on the same database, which utilise the extracted features as the input to their models. Therefore, this paper provides evidence favouring the use of end-to-end networks in emotion recognition through speech, which do not rely on pre-processed audio data. However, it would be interesting to see whether this novel model takes longer time to train, which the authors failed to mention.

End-to-End Multimodal Emotion Recognition using Deep Neural Networks

Tzirakis et al. (2017) explore a deep learning end-to-end approach to emotion recognition while considering the audio and visual modality. Similarly, they focus on raw data as the input to their models, rather than extracted features, however not only using audio signals but visual as well. Their network relies on a CNN to extract the audio features and ResNet-50 network (He et al., 2016) for visual features. Following, fusion of the multimodal data takes place and a LSTM network is responsible for emotion detection consisting of arousal and valence prediction.

Using the RECOLA dataset, which in addition to speech, contains visual data, the model's performance was assessed. With regards to the audio modality, the model outperformed all competitors when predicting arousal, but did not manage to achieve similar results for valence. On the other hand, visual modality alone outperformed all other models using the dataset on valence and most on arousal. When considering the combination of modalities, this model was significantly better at predicting valence but not necessarily arousal. These results provide mixed evidence in support of the end-to-end network with raw signal as input. On average, the authors' model performs better than competition, however there are cases where the other models which use extracted feature have better results, suggesting that both methods are worth considering.

Evaluating deep learning architectures for Speech Emotion Recognition

Fayek et al. (2017) take a rather different approach, utilising speech emotion recognition as a way to test various deep-learning architectures. More specifically, the overall model is of end-to-end nature passing input through deep multi-layered neural network to obtain class probabilities, with the network varying in architecture. The authors experimented with various combinations of layers within the architecture, including different sizes of spatial and temporal convolutional layers and fully connected layers with batch normalisation, Rectified Linear Units (ReLU's) and dropout.

The architectures were compared based on their performance on the IEMOCAP dataset. The best performing architecture consisted of two spatial convolutional layers followed by two fully connected layers. This model achieved better accuracy in comparison to a LSTM-RNN and a 5 layer fully connected architectures designed by the authors. Additionally, the authors compared the model's performance against prior work with IEMOCAP and at the time achieved state-of-the-art results. However, when the other models utilised additional modalities, this did not hold. Additionally, it would be interesting to see their chosen model in comparison to a combination of a CNN and LSTM, which is often employed in the other papers. Overall, this paper provides and interesting overview of various model architectures that can be used for speech emotion recognition.

Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning

Li et al. (2019) proposed a new method for emotion recognition using speech. To specify, their method involves end-to-end architecture with two interesting features, multitask learning and self-attention. Multitask learning refers to training the model to perform more than one task at once, which should allow the model to share the information from one task to benefit the other. Specifically, within this paper gender classification of the speaker is employed as a second task, as the authors hypothesise that emotion recognition and gender classification share similar features. The proposed model consists of a spectrogram used for feature extraction, and a self-attentional CNN-BLSTM. The interesting element in this model is the self-attention network which purpose is to gather information from hidden states of the BiLSTM to transform the speech data into a vector of a certain length. The output of this network is then fed into the two output layers, emotion and gender.

To assess the model's performance, the IEMOCAP database is used to train and test the model. The authors provide the performance statistics for the full model as well as separately for self-attention and multitask learning. Overall multitask learning is significantly better when compared to the self-attention, but the combination of both results outperforms any previous work on the IEMOCAP dataset significantly. Due to these results, the addition of such mechanisms when devising the emotion recognition model should be considered.

Summary

When considering whether to include the audio modality in the emotion recognition model, consulting the above described papers was quite beneficial as the speech modality also seems to carry importance in detecting emotions. Furthermore, if the modality is included one should consider what architecture should be chosen for processing speech as well as the data pre-processing. Based on the research portrayed, a combination of a CNN and a LSTM network seems the most appropriate architecture to use. Additionally, often the comparison is made between whether the input to the network should include extracted features from the audio or the raw

audio signal. Based on the results of the papers, it can be concluded that utilising raw data comprises the state-of-the-art in terms of performance. Additionally, extracting features from audio can be time consuming due to having to select the appropriate extraction method, as different datasets could benefit more from certain feature sets. Overall, it would be beneficial to experiment with the end-to-end network architectures relying on unprocessed data.

2.4.3 Prior Work on CMU-MOSEI

Multi-task learning for multi-modal emotion recognition and sentiment analysis

Akhtar et al. (2019) approach emotion recognition on the CMU-MOSEI using all three provided modalities, as well as multitask approach incorporating sentiment analysis. Their proposed model explores the contextual inter-modal attention mechanism, in order to utilise both the context and the attention mechanism. This is done by incorporating bi-directional Gated Recurrent Units to obtain context and followed by pair-wise inter-modal attention mechanism to benefit from the connections between the modalities. Subsequently, the output is concatenated and fed into the output layers for both tasks. With regards to emotion recognition, the authors chose to approach this a multi-label classification problem predicting the six basic emotions as well as no emotion, which in CMU-MOSEI is regarded as when an utterance has rating of 0 on all emotions. In terms of the loss function, the output is first fed through a sigmoid layer followed by binary cross entropy per each emotion.

One of their aims is to evaluate the different combinations of modalities and their effect on the performance of the model. From the results, one can see that the textual modality outperforms the other two. However, a combination of any two is still better and the best results are achieved with all three modalities, but only marginally. When considering two modalities, the best performing combination is text and video, followed by text and audio, with audio and video performing the worst. Furthermore, multitask learning outperforms single task. When comparing the model's performance with prior works on CMU-MOSEI, this multitask model achieves state-of-the-art on most emotions. However, when considering happiness, the model performs worse than the others.

Multi-modal sequence fusion via recursive attention for emotion recognition

Beard et al. (2018) propose a network utilising recursive multi-attention and a shared memory for multimodal emotion recognition. This approach features extracted features from all of the data as an input to their newly proposed Recursive Recurrent Neural Network which should be similar to a normal RNN, however incorporating recursive hidden state. Overall, the network is able to maintain an external memory useful for operations with the incoming input.

In comparison to other papers, the authors chose to approach this task as a regression rather than classification task. The results seem to be very promising, but perhaps due to regression being a less popular approach for this dataset, the authors do not mention any comparison to previous works, which limits the extent to which one can evaluate the efficacy of this model.

A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis

Delbrouck et al. (2020) describe a different approach to emotion recognition on CMU-MOSEI, utilising Transformer architecture, called the Transformed-based joint-encoding . The transformer architecture is chosen specifically for parallelisation suitability. Authors compose two models, one for single modality and a multimodal one combining all three modalities. The models are composed of transformer blocks, with the multimodal one benefiting from shared attention and a final classification layer. The model relies on pre-extracted features from data. The task is approached as multi-label classification with 6 emotions to be predicted.

When evaluating the model, the statistics are slightly misleading. Firstly, accuracy is used as a metric to evaluate the performance but due to the high per class imbalance, it is not a reliable metric to employ. This is especially visible once other metrics are consulted. Secondly, as other works used different metrics, the authors do not compare their results to previous work, and therefore their results cannot easily be set into context.

Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis

Chauhan et al. (2019) relied on RNN approach, the Context-aware Interactive Attention, including multitask learning of emotion and sentiment. Their model utilised the interaction between any two modalities using a similar structure to an auto-encoder. This is followed by using Bi-directional Gated Recurrent Unit, which is concerned about the sequential aspect of utterance. Sequentially, a Context-aware Attention Module is employed to recognise patterns between sequential utterances. In the end, the model predicts the intensities of the emotions, adopting a multi-label classification approach. Similarly, to other works, a sigmoid layer followed by binary cross entropy is employed in training.

When judging the effect of different modalities on the performance, we can see a familiar pattern. Text outperforms the other two, followed by audio, and a combination of text and video outperforms the rest, followed by text and audio. However, the combination of all three seems to be the preferred option. When comparing the multimodal model to any prior work, the results are significantly better, resulting in state-of-the-art performance.

Summary

Analysing prior work on the selected CMU-MOSEI dataset was extensively informative. It allowed us to decide on the multi-label classification approach, as due to its higher prevalence it will allow for better comparison of results achieved. Therefore, we will be able to set our model into context in terms of how reliably it can classify emotions. Furthermore, similar to the mentioned papers utilising multilabel classification, we can follow the notion of calculating loss through per class binary cross entropy. Moreover, as the approaches and devised networks are quite varied, there does not necessarily seem to be one best network architecture to follow. It is worth to mention that all of the papers use some level of feature extraction. It should be interesting to compare their performance with a model utilising raw data as input, and due to this and other reasons portrayed in the previous sections, this is likely how we will approach this task. Similarly, none of the described models utilise BERT for the text modality. As BERT is considered the state-of-the-art in NLP, if we choose to incorporate it within our model, another comparison can be made. Even more interesting is, the CMU-MOSEI emotion labels consist of intensity ratings, and none of the papers seems to take this into account. Rather, most treat any intensity greater than 0 as the presence of an emotion. It could be possibly worthwhile to somehow incorporate these intensities in our training as they may carry a level of importance and potentially improve the model. Additionally, it is apparent that the best results are achieved using all three modalities. However, as discussed before the visual modality would not be usable for the purpose of the model. Therefore, when comparing audio vs text vs their combination, to achieve the best results we should incorporate both modalities based on the discussed research.

2.5 Ethical and Professional Considerations

It is important to consider the possible consequences of this project, more specifically any legal, social or ethical issues that may arise. The main focus of this project with regards to techniques used, involves ML and there are various implications to consider when utilising ML within a project. The Ethics Checklist is provided in Appendix D.

Firstly, ML usually includes long training times on a large dataset, meaning that often a great amount of computer resources needs to be allocated. As a consequence, if a shared server is used for training, this may limit the amount of resources that can be allocated to others, whether in terms of storage or computing power. To address this, the training process will be made efficient, in order to not block out access or shared resources for other potential researchers and users.

Secondly, we are aware of the potential bias that can be produced by ML, which may stem from using certain datasets, which lack a diverse representation. Additionally, in terms of emotion recognition, even though certain emotions are deemed as universal, there still could be a cultural aspect affecting them or their expression.

Therefore, based on the dataset chosen, the trained model may discriminate against certain unrepresented parts of the population and therefore suffer in terms of external validity. This issue is not easily addressed, as most large datasets in English language are not necessarily diverse. However, choosing an appropriate dataset such as CMU-MOSEI, which is large enough and diverse in terms of speakers and topics, as well as consisting of real-life data, could to some extent mitigate this.

Thirdly, when considering the virtual therapist aspect of the project, ML was not utilised when guiding the dialogue between the therapist and the user. This choice was made specifically, to avoid any unpredictable behaviour from the therapist. As such behaviour could negatively impact the possible vulnerable population, people with mental disorders, who should benefit the most from SAT. Therefore, the responses of the therapist are rather based on a professionally devised script.

Finally, we would like to mention that none of the data used within this project has been modified to produce satisfactory but biased results. Similarly, all reported results are real and an effort was made to not report only the positive results, but rather to provide a fair evaluation and set the results in wider context by including various statistics. For future work purposes, we would also like to note that using the emotion recognition model on any human subjects, will require ethical approval from a professional or academic body such as Imperial College London.

Chapter 3

Emotion Recognition Model

To summarise some of the design choices, we are interested in implementing an end-to-end network architecture, with raw data as input, consisting of BERT for processing text and a combination of CNN and LSTM for processing audio. Moreover, the model should detect any of the six basic emotions, therefore we would like to employ multi-label classification.

Fortunately, we were able to obtain an already existing model in line with our design choices and in addition an emotion recognition framework from James Tavernor, who has previously worked within the group project. Tavernor (2020) has devised and implemented an emotion recognition model and was able to achieve state-of-the-art results using the IEMOCAP database. Originally, it was intended for their model to be used within the VR SAT application, however due to the chosen database, the model can reliably predict these emotions: happiness, sadness, anger, neutral, frustration and excitement. As mentioned before, we are interested in being able to predict similar emotions but specifically including fear due to its relation to anxiety. Furthermore, as emotions are not binary and often co-occur, to achieve greater external validity, instead of multiclass classification, which predicts only one emotion at a time as done by Tavernor, we will engage in multi-label classification. Consequentially, combining these possible improvements with the state-of-the-art model should result in good performance and especially better external validity, which is important due to its inclusion in the VR SAT application.

Additionally, it is possible that in future implementations and work on the VR SAT application, in order to improve the emotion recognition model even further, one would wish to combine the training on both of the databases. For this reason, in addition to time efficiency, the emotion recognition model will be implemented using Tavernor's framework. This will result in the same code base for both of our projects, which should make it easier and accessible to improve upon it in the future. As it is already challenging to understand one's code, having to understand both in order to combine them or build up on them would be unnecessarily hard. As Tavernor has already implemented a well-done emotion recognition framework, expanding upon it to accommodate for additional dataset and multi-label classification is advantageous.

3.1 Model

As mentioned, the model architecture used is provided by Tavernor (2020) and so in this section its architecture and any changes made will be outlined. Tavernor's model utilises two modalities, speech and text in order to predict one of the mentioned emotions. For the audio modality, Tavernor has adapted an already existing model by Rizos et al. (2020) resulting in two sub-models. Firstly, the CNN model is responsible for processing the raw audio data, similarly to previously mentioned research. Secondly, the BiLSTM model processes the extracted data from the CNN model. The BiLSTM model consists of LSTM stack and an attention pooling layer, followed by a liner layer. With regards to the text modality, the BERT architecture is used, more specifically BERT-large, which consists of 14 transformer blocks and a hidden size of 1024, with 340 million parameters (Devlin et al., 2018). Tavernor has chosen BERT-large, as in comparison to other architectures it performed the best. The text data needs to be tokenised and encoded into transcript ID's and an attention mask to be inputted to the model. Following, the BERT architecture outputs the [CLS] token which is then processed by a ReLU layer. After both modalities are processed by the above described methods, they are concatenated to fuse the outputs together. Tavernor refers to this as mid fusion, which was chosen due to its positive effect on performance. Furthermore, a soft-max layer is employed to calculate the modality attention weights which are then combined with the output using multiply and sum operations. Finally, the model employs linear layers per each task predicted, which in Tavernor's case includes arousal, valence, dominance and emotion. Additionally, before the linear layer is applied for emotion prediction, Tavernor uses a custom LSTM for modelling speaker level memory.

Overall, three minor changes were made to this model to accommodate it for our purposes. Firstly, we are not interested in predicting arousal, valence and dominance, nor does the CMU-MOSEI provide such data. Tavernor has already included a way to not employ these layers of the model through a variable. Secondly, the speaker level memory was specifically designed for the IEMOCAP dataset as the dataset consists of 5 conversations between 10 actors. Including the speaker level memory allows to remember the emotional state of an utterance and therefore assess its effect on the sequential utterances spoken by a single speaker. However, due to the different specifics of our dataset we are not interested in this and it was removed, resulting in only the linear layer being applied to the output. Lastly, the output size of the linear layer was adapted to the number of emotions we would like to predict, which again was simple due to Tavernor's model design.

3.2 Framework

Tavernor's framework (2020) consists of several scripts responsible for various parts of implementing an emotion recognition model. It includes various training loops that Tavernor has used for training their model, as well as model architectures, custom datasets and other scripts enhancing functionality or ease of use. As much as

the framework has accelerated the progress with regards to training the model, there were several challenges to overcome. Firstly, the framework does not have any documentation to explain its functionalities as well as only small number of comments. As a result, it took some time to fully understand its capabilities, through extensive research, but Tavernor has kindly pointed us in the right direction when approached. Secondly, the framework even though build to be used to expand upon Tavernor's work, is very specific to the dataset and techniques Tavernor used. To illustrate, there was no easy way to adapt the certain scripts such as abstract methods etc., although Tavernor is working to improve this. Therefore, parts of the framework are often utilised as a template, when creating new content specific to this project. The advantage of this, is that it took significantly less time to get into the training process as well as previously discussed, it will be much easier to understand for anyone wishing to work on this further. More specifically, this method was used when creating a custom dataset and a way to load it for training and then for implementing a training script. For other parts, the framework was sufficient to use as is or with very minor changes. Overall, acquiring the access to the framework has significantly helped in terms of reducing the time taken as well as gaining further understanding of the relevant methods used.

3.3 Database

To easily obtain the CMU-MOSEI database the authors provide an SDK, called the CMU-multimodal SDK (Zadeh et al., 2018), which allows for loading and working with various datasets. The datasets are loaded in computational sequences and with regards to CMU-MOSEI, the data consists of already pre-extracted features. Due to us wanting to utilise the end-to-end network capabilities, using raw data as input to the model, this was not desirable. Therefore, instead of working with the SDK, it was necessary to seek out the unprocessed data. The authors have made such data available to download through a link provided in their description of the SDK. One should note that it was not possible to only download the audio and text data but all modalities, consequentially the visual data was discarded. The audio data consisted of audio recordings in the wav format, with each recording representing multiple utterances and therefore had to be separated. The text data was provided in the form of txt file, per every recording one text file was provided with time-stamped utterances and their transcriptions. With regards to the emotional labels, a computational sequence was used to hold the data in place, accessible by a key matching to the names of the audio recordings and text files. Following, two arrays were held by the sequence, one denoting the timestamps associated with the utterances and the other representing the emotional labels per each timestamp as an array of 6 numbers, ranging from 0 to 3. After the database was obtained, the descriptive statistics were calculated and are shown in Table 3.1.

Total number of audio files	3225
Total number of utterances	22860
Average audio length [utterances]	7.09
Average utterance length [seconds]	7.28
Average intensity (excluding 0-intensities)	0.53
Average intensity (including 0-intensities)	0.17

Table 3.1: Descriptive statistics of CMU-MOSEI dataset.

Furthermore, the distribution of the classes can be seen in Figure 3.1, accompanied by an exact count in Table 3.2. As mentioned before, happiness is by far the largest class, with sadness, anger, disgust and other/no emotion being moderately represented and fear and surprise unrepresented. Consequently, the dataset is quite imbalanced, and we can expect better results for the larger classes. Therefore, it would be worth to consider various strategies to account for the class imbalance.

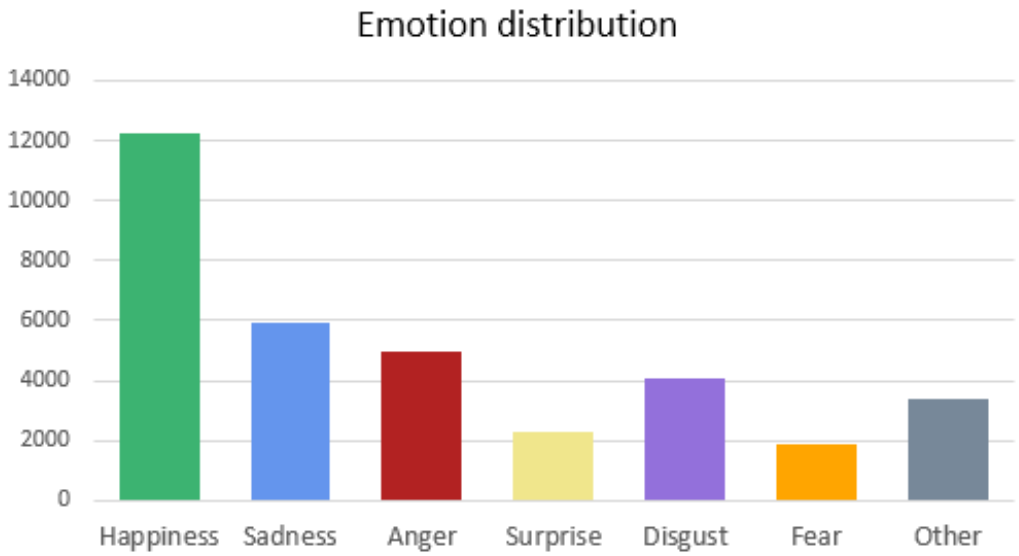

Figure 3.1: Distribution of emotion classes in CMU-MOSEI dataset.

Emotion	Total
Happiness	12245
Sadness	5918
Anger	4935
Surprise	2286
Disgust	4041
Fear	1892
Other	3395

Table 3.2: Total number of samples per emotion class.

3.3.1 Train/Valid/Test Split

For the purposes of training the model, the data is divided into three folds. Firstly, a train fold to provide the data for the model to train on. Secondly, the validation fold to be used for finetuning the model's parameters and thresholds. And lastly, the test fold for evaluation of the model's performance on unseen data. In order to determine these splits, we have adapted the standard splits as provided by the authors (Zadeh et al, 2018) in order to assure speaker independency. The ratio is approximately 70/10/20 % for train, valid and test set, respectively. Additionally, the folds consist of a similar distribution of the emotional classes as can be seen in Tables 3.3, 3.4, 3.5.

Emotion	Total	Ratio
Happiness	8735	0.35
Sadness	4269	0.17
Anger	3526	0.14
Surprise	1642	0.07
Disgust	2955	0.12
Fear	1331	0.05
Other	2391	0.10

Table 3.3: Distribution of training set.

Emotion	Total	Ratio
Happiness	1005	0.36
Sadness	520	0.18
Anger	338	0.12
Surprise	203	0.07
Disgust	281	0.10
Fear	176	0.06
Other	302	0.11

Table 3.4: Distribution of validation set.

Emotion	Total	Ratio
Happiness	2505	0.36
Sadness	1129	0.16
Anger	1071	0.15
Surprise	441	0.06
Disgust	805	0.11
Fear	385	0.06
Other	702	0.10

Table 3.5: Distribution of testing set.

3.3.2 Data Pre-processing

As discussed before, we opted out of using extracted features and would rather like to process the raw signal from the audio data. Therefore, the data pre-processing consisted of the following steps. Firstly, the data was loaded into a custom dataset, consisting of multiple dictionaries storing the emotional labelling, in the form of original intensity and multi-hot vectors, and transcripts. The audio data was not stored in the dataset structure as this would take a large amount of memory but was rather loaded from the associated files, whenever required. After the initial

load, the audio data is resampled to the same sampling rate. Furthermore, based on the descriptive statistics, calculated on the train set of the audio recordings, more specifically mean and standard deviation, the audio data is additionally normalised when loaded. The normalisation refers to transforming the data to z-scores based on the formula $z = \frac{(x-\mu)}{\sigma}$. As all emotion classes are represented enough to some extent in the dataset, no data pruning took place and we were able to utilise its whole size.

3.4 Training

After the data was pre-processed and the custom dataset was ready to load the data, training implementation began. As mentioned above, using certain parts of Tavernor's (2020) framework as a template, reduced the time taken significantly. We have chosen to use Pytorch, a Python machine learning framework, as Tavernor's model and framework were built on it and due to previous experience with it. First step in training was to create an object responsible for loading the data and dividing it into batches. This could have been easily done using Tavernor's script however, some changes needed to be made.

As Tavernor has only used a train and test set, for training purposes instead of loading these, we were interested in loading the train and validation set. Furthermore, as the IEMOCAP data Tavernor used this script on, was in a different format, a custom object was used to sample the data. For the purpose of this project, we were interested in a random sampler, that would shuffle the data as to not feed it into the model in the same order. Following, the training loop needed to be adapted by removing any references to arousal, dominance and valence training, as well as any functionalities regarding the conversation aspect of IEMOCAP and some other minor changes. Tavernor has implemented a saving functionality which allows to save the model every n-selected iteration, after running an evaluation and calculating various metrics, as well as load an already saved model and continue training on it. This functionality was utilised often as it also allowed to consider the effect of each additional epoch on training, by inspecting the various associated metrics which get saved in a text file. Regarding the metrics calculated, they were slightly adapted to fit the multi-label nature of our work. Following, an appropriate loss function was selected by consulting previous research and various Pytorch options. The loss function BCEWithLogitsLoss was chosen as it incorporates a sigmoid layer and binary cross entropy (BCE) per each emotion, which was often what was utilised in prior work on CMU-MOSEI. Initially, as input to the loss function we have chosen multi-hot encoded target emotions, which were constructed by transforming the intensities to 1 whenever > 0 and 0 otherwise. The loss function could also include weights based on the classes however this was not employed in the initial stage. A final change, regarding the validation script, the predicted and target emotions were processed differently due to the multi-label aspect of our training.

After the above changes were carried out, the training could have begun. Initially, an attempt was made to train the model on personal computer trying training on CPU or GPU using CUDA. With regards to CPU training, even if the RAM limit hadn't been reached, it would still take an incredibly long time to train on epoch, approximately 17 hours. When considering GPU, the training time sped up significantly, however out of memory errors were encountered, due to the large size of the model. Therefore, this has resulted in training the model on the Imperial College London provided GPU cluster with more powerful GPU's. However, even then some memory errors were encountered.

In attempt to fix them, it was noticed that Tavernor's evaluation script is not memory efficient and over time consumes more memory. This is not a normal occurrence and usually results from storing variables without detaching the gradients. After conducting additional research on how this could be fixed, the issue was resolved, as indeed certain variables and gradients were unnecessarily stored. Since then, no memory errors were encountered, as well as now the memory usage was stable throughout in that portion of the code. Additional strategy, which was also employed by Tavernor, was to only load and calculate the loss on 2 samples at a time and then to update the parameters of the model based on the selected batch size, which was successful in reducing the amount of memory errors during training. However, still a small number of memory errors appeared caused by the training script. The interesting part was that a small number of certain samples, likely the larger ones, were responsible for the errors. Therefore, the script was adapted by ignoring any samples that would cause out of memory error. This had solved the issue and the first training began properly. Of course, after the training had finished, it was confirmed that this solution does not skip too many samples, which would be undesirable, but in the end only a small number of samples was skipped. With the use of the GPU cluster, which limits the user to one GPU, the training time has significantly reduced to one epoch taking approximately 5-7 hours to complete, together with the validation of the epoch which take approximately 30 minutes.

After successfully training a model, the first results based on the validation dataset were decent and to increase the performance, the following techniques were explored. To assess the performance, the metrics calculated on the validation set were consulted. More specifically, the greatest focus was set on macro and micro averaged AUROC as it is independent of the thresholds set for classifying emotions. Whereas the other metrics such as accuracy and F1 scores are calculated based on the set threshold, which was yet to be selected per each model during evaluation to maximise the performance on these metrics. Additionally, confusion matrices were consulted, with analysing each matrix per emotion.

The model includes a parameter hidden size which allows to change the input, hidden and output sizes of certain layers in the model. In order to inspect the effect of this parameter on training time and the model's performance, three variations of the model were trained. This involved training the model with the parameter set to

128, 256, 512 and 1024 respectively. After consulting the various metrics there was no significant difference in the overall performance of the model between 128, 256 and 512 hidden size. However, a marginal increase in performance was noted for the 1024 one. As this parameter did not seem to affect the training time to a large extent, the 1024 model was selected as the best performing model.

Initially, the threshold that had to be crossed in order to detect the presence of an emotion was set to 0.5 and this threshold was utilised for all emotions. Afterwards, this number was varied to assess the effect of the threshold on separating the positive and negative classes. However, it became clear that a singular threshold does not allow to maximise emotion detection for all classes in the same way. Lowering the threshold was needed for some of the less represented classes, whereas for happiness often a greater threshold could be set. Therefore, the strategy of setting custom thresholds per emotion was adopted, to maximise the model performance. These thresholds ranged variously and to select the best threshold the confusion matrix per each emotion was considered, as it provided the most information.

Soon it was apparent, that the imbalance in this dataset has effect on the performance of the model. Mostly, as better performance was registered when recognising happiness, the most represented emotion, whereas for the other emotions the performance was often lower, especially for surprise and fear. To mitigate this, the initial and the simplest strategy was to assign weights to the classes and to utilise these weights during loss calculation. This technique should cause the model being penalised more strongly for the less represented emotions and therefore should result in a similar performance as if the model trained on a balanced dataset. The loss function chosen, did in fact provide such functionality and the documentation was consulted to see how to format the weights.

More specifically, the binary cross entropy loss requires the class weights in the form of weights for the positive classes. Meaning that we consider the positive class as the selected emotion and all the other samples act as a negative class. These weights are calculated per class, in the form of negative samples / positive samples. After including the weights in the loss calculation, their effect on the model's performance was interestingly insignificant. For certain classes the performance was marginally better and for other marginally worse, resulting in no overall performance increase. Following, an additional instruction included in the loss documentation was noticed, which mentioned that weights > 1 should increase the recall metrics, while weights < 1 should increase precision. As high recall was achieved quite easily by the model, while the precision seemed to suffer the most, the strategy was to include the < 1 weights. To format the weights, a sigmoid layer was used on them to distribute them between 0 and 1. After these weights were added, again no significant effect was observed in the performance of the model.

The initial experiments produced favourable and competitive performance of the model. To increase this performance even further, an additional technique was adapted, instead of using the ground truth, the model can operate under uncertainty (Rizos & Schuller, 2020). More specifically, instead of using the multi-hot encoded emotional intensities as the target emotion when calculating loss, we would utilise soft labels, which represent the distribution of the emotions. The idea is, that the model should be corrected more strongly if a high-intensity emotion is misclassified. Whereas if the emotion is of low intensity which can also be a result of disagreement between the raters, the model will be penalised less.

To acquire the soft labels, the given intensities ranging from 0 to 3, need to be transformed to range from 0 to 1 as this is the only accepted input of the selected loss function. Two methods were used, firstly using a sigmoid function on the intensities to transform them and secondly dividing the intensities by the highest possible intensity. The first method is more arbitrary but resulted in an improvement over using the multi-hot encoded labelling. However, the second method was chosen, as it is more representative of the actual distribution and it outperformed the first method. Specifically, the greatest improvement was in detecting happiness, whereas surprise did not benefit from this method. Additionally, both weight types were tried again and using the < 1 weights, the performance has increased marginally for certain classes. As there was no negative effect of including the weights, they were kept.

Additionally, it is worth noting that in the first implementation of soft labels, the model was trained only using the six basic emotions, whereas the previous experiments had 7 emotion, including the other/no emotion class. The reasoning was, that other/no emotion had no intensity associated with it through rating. However, for comparison purposes we have decided to train another model, using soft labels with the inclusion of the 7th class. In order to assign the intensity to the other/no emotion, an average intensity across the training dataset was calculated. Interestingly, the performance of the model reduced significantly after the inclusion of the other/no emotion class.

After experimenting with soft labels, the performance of the model, specifically regarding happiness was impressive. Therefore, it was decided to use a different technique to combat the class imbalance, as using the weights did not help significantly. This involved adapting the loss function from the current BCE loss to focal loss. Focal loss is a relatively novel loss which was devised in order to address class imbalance (Lin et al., 2017). To do so, it adapts standard cross entropy, to give more weight to the difficult samples and less to the well classified ones. Furthermore, there is evidence to suggest that focal loss should outperform standard cross entropy (Mukhoti et al., 2020) Therefore, including such loss could potentially improve the performance of our model, by updating parameters in a way that gives more focus to the difficult-to-classify samples. Pytorch does not natively include focal loss and therefore a custom loss function had to be used. After conducting some research, a fitting online implementation of focal loss was found and used (Qin, 2018).

There are two hyperparameters that can be adjusted within the focal loss function, alpha and gamma. Alpha is selected to weigh the samples, ranging from 0 to 1. If a sample belongs to the positive class, it is weighted by alpha and if negative then by 1 – alpha. Gamma parameter determines the focus on the easy/difficult-to-classify samples. With a larger gamma more weight will be assigned to the difficult samples and less to the easy ones. The most common values for alpha and gamma is 1 and 2 respectively, however those parameters can be tuned dependent on the problem. Initially, the loss function was replaced to feature focal loss using the common alpha and gamma values, which again reduced the performance significantly. Therefore, several other combinations of the parameters were tried, with the values of 0.75 and 4 for alpha and gamma respectively resulted in the best performance. However, it is worth noting that due to limited time not all options were exhausted and therefore further hyperparameter tuning would be beneficial.

Due to the long training time, not many options were explored in terms of hyperparameter tuning. In terms of the initial learning rate parameter, the options ranged from 0.001 to 0.00001, with a smaller learning rate often correlating with a better performance. Furthermore, the Adaptive Moment Estimation (ADAM) optimiser is used, which features adaptive learning rates for the different parameters in the model. During the initial stages of the training, specifically when trying to combat the memory errors, the Stochastic Gradient Descent (SGD) was used as well. However, a significantly lower performance was recorded with SGD when comparing to ADAM. Similarly, not all options were exhausted when considering the batch size parameter. Due to memory issues only 2 samples could be loaded at once with the loss being calculated for each two samples. However, the parameter update was dependent on the batch size and the possible values that were tried ranged from 2 to 64, with 32 resulting in the best performance. The effect of the number of epochs was tracked across the whole training as the model was usually trained for multiple epochs whenever a parameter was tuned. This was due to certain parameter possibly affecting the number of epochs needed to achieve the best performance. Most of the time, a smaller number of epochs was preferred, as it seemed that with larger epoch number the model tended to overfit the training data and perform worse on the validation set.

3.5 Final Models

The following models were selected to be evaluated on the test set, in order to compare their performance and assess which model should be preferred. Firstly, the hard label model which acted as a baseline for the other experiments, which we will refer to as the baseline model. Secondly, the soft label model, where the labels were calculated by dividing the intensities to achieve probabilistic distribution, including the < 1 weights but not the other/no emotion class, which we will call SoftW model. Lastly, SoftF model, which is similar to the previous one, but instead of using weights to combat the class imbalance, the model used a different loss function, the focal loss. The different parameters of these models can be seen in Tables 3.6, 3.7, 3.8.

Parameters	Baseline model
Hidden size	1024
Thresholds	[0.4, 0.25, 0.2, 0.1, 0.25, 0.1, 0.1]
Loss function	BCE
Labels	Hard
Number of classes	7
Learning rate	0.0001
Batch size	32
Number of epochs	3
Optimiser	ADAM

Table 3.6: Parameter specification for the Baseline model.
Thresholds are displayed in the order of happiness, sadness, anger, surprise, disgust, fear and other.

Parameters	SoftW model
Hidden size	1024
Thresholds	[0.06, 0.03, 0.03, 0.015, 0.02, 0.017]
Loss function	BCE + weights
Labels	Soft
Number of classes	6
Learning rate	0.0001
Batch size	32
Number of epochs	5
Optimiser	ADAM

Table 3.7: Parameter specification for the SoftW model.
Thresholds are displayed in the order of happiness, sadness, anger, surprise, disgust and fear.

Parameters	SoftF model
Hidden size	1024
Thresholds	[0.25, 0.195, 0.19, 0.15, 0.15, 0.16]
Loss function	Focal Loss
Labels	Soft
Number of classes	6
Learning rate	0.0001
Batch size	32
Number of epochs	4
Optimiser	ADAM

Table 3.8: Parameter specification for the SoftF model.
Thresholds are displayed in the order of happiness, sadness, anger, surprise, disgust and fear.

3.6 Challenges

When considering the whole process of coming up with the final emotion recognition model, various challenges were met. Overall, this process was spread out almost from the beginning of the project till its completion. Firstly, the research portion took a long time, which consisted of gaining a level of understanding of the various state-of-the-art techniques used in emotion recognition. However, the time spent on research, translated to less time spent on figuring out the various other implementation steps due to better understanding. Secondly, the time it would take to understand fully the functioning of Tavernor's (2020) framework was underestimated, which was also prolonged due to the lack of documentation. Similar to the above, once we were confident enough about the separate components of the framework and their integration, using it was considerably easier and the focus could have shifted on implementing the emotion recognition model. Thirdly, we were not aware of the time it would take to train a model using such a large dataset. As a result, tuning hyperparameters had to be done manually, which was quite slow and lesser variety of different parameters could have been tried before the project ran to its completion. Additionally, a lot of time was spent on fixing the initial memory errors, which were resulting from the combination of large model and the dataset. Lastly, to be able to train such model within an acceptable time frame we had to resort to training on a GPU cluster, which was initially time consuming. However, once the process was understood, training on the cluster decreased the training time significantly, in comparison to a personal computer.

Chapter 4

Virtual Therapist Integration

4.1 SAT Scenario

The SAT scenario to be implemented, was obtained from Prof. Abbas Edalat, with some adaptations made by Neophytos Polydorou, who is responsible for developing the virtual reality platform for the application (Polydorou, 2020). The scenario consists of four stages, Introduction to Self-attachment therapy, Connecting with the childhood self, Falling in love with the childhood self and Developmental exercises for the childhood self. The scenario will be outlined below, with the full scenario along with the descriptions of the associated virtual environment, as devised by Polydorou, displayed in Appendix A.

In the first stage, the therapist introduces themselves, the platform and SAT to the user. In the second stage, the user is introduced to their childhood self and asked to interact with it to start developing a bond. In the third stage, the user is asked to sing to the child to further deepen their emotional bond in order to be ready for the following stage of the therapy. Within these three stages, emotion recognition should be conducted mainly to track user's emotional state. In the fourth stage, which is the most important stage of the therapy, the user follows six exercises, focused on reparenting the childhood self as well learning proper emotion regulation.

The first exercise, Sessions for processing the painful past, consists of the user being asked to remember a traumatic episode. Based on the user's emotional reaction to this, which is likely to be a negative emotion such as sadness, anger, fear or perhaps disgust, this emotion will be transferred to the child avatar. Afterwards, the user is tasked to parent the child by reassuring, embracing or cuddling the child, after which the avatar should display happiness. Within the second exercise, Sessions to process the current negative emotions, the user is asked to recall most recent negative events to elicit the negative emotions associated with them. Through emotion recognition, the detected emotions are transferred to the child avatar. Similar to the first exercise, the user should interact and parent the child, after which the child's emotions should change eventually to happiness. The third exercise, Protocols for creating zest for life, involves the user singing to the child in order to change its

emotional state from neutral to happy. Through this interaction the user is taught the effect of singing on eliciting joy, which should transfer to their real life. In the fourth exercise, Getting over the negative emotions, the user will be shown an image of the Gestalt vase and will be asked to focus on its certain features, based on the therapist guidance. The fifth exercise, Socialising protocols for the childhood self, consists of a self-massage done by the user and a confirmation of completing the previous exercises. In the last exercise, Creating a more optimal internal working model, the user is introduced to the idea of internal working model, represented as a house built piece by piece based on exercise completion. The user's emotional state is detected, through emotion recognition and their progress is tracked to see the effect of the therapy on user's mood.

In addition to the scripted scenario described above, after discussing the possible options we have agreed to include an additional response from the therapist. More specifically, the option was added after emotion recognition takes place and the therapist asks the user to confirm the detected emotion. Whenever the user disapproves, the therapist asks the user to select their preferred emotion. This allows for greater precision when projecting the emotions of the user to the child avatar.

4.2 Design

Due to the nature of this project, one of the most important things to consider, is to allow for easy integration with the VR application. As mentioned before, the virtual environment is built in Unity, which is a real-time 3D development platform, utilising C# scripts. The main reason for choosing FAtiMA engine, is that it can utilised as a C# library, which should be easily integrated with Unity. In addition, FAtiMA provides a tool for creation of dialogue or action-based scenarios. Therefore, rather than implementing the whole dialogue using a script, the decision was made to implement the scenario with the FAtiMA-provided tool. However, this tool does not provide the option to allow the user to say unscripted sentences, rather the user is bound to select one from number of options. This is undesirable for our application and therefore it had to be adapted in the C# implementation. Additionally, the therapist lines had to be customised, whenever the scenario requires the therapist to react to the current emotion of the user, e.g. I see that you are [current emotion]. Furthermore, for debugging purposes, a console-based application was built to check whether the dialogue was correctly implemented. With regards to integrating the emotion recognition model, as the model is created in Python, we had to consider how the C# code will interact within this model. As well as, whether such integration is possible on the Oculus Quest device, which is based on Android operating system and there may be limitations to what can be run on it. Therefore, the possible integration had to be discussed with Polydorou, to provide them with a product that is easily integrable.

4.3 FAtiMA Implementation

The FAtiMA Integrated Authoring tool provides the capabilities of creating various scenarios including multiple characters, each with different existing beliefs, a common word state as well as a dialogue editor. To create the SAT scenario, the first step was to build the applicable characters. As the only characters which engage in the dialogue in the SAT scenario are the user and the virtual therapist, these were created first, called player and therapist respectively, followed by forming their decision making. As the scenario in SAT is quite straightforward and scripted, the only applicable actions that the characters may perform consist of speech. However, if one would later plan to expand on the VR SAT application, FAtiMA allows for additional actions to be added, as well as alternative dialogues from which the virtual therapist may choose from either based on priority or other variables.

To form the decision-making process of the character, an action rule and relevant conditions need to be added in the tool. The action rule consists of the action name, which in our scenario is always the action speak, consisting of the following variables, current state, next state, meaning and style, displayed as Speak([cs], [ns], [m], [s]). The current and next state refer to an arbitrary point in time, in which the parts of the dialogue reside. Only the options whose current state matches the state of the world can be selected. The meaning and style variables are again completely arbitrary and can be used to further specify which dialogue option should be chosen. In addition to the name of the action, a target of the action is selected and the priority of the action. As our scenario does not allow for any agency in terms of actions the priority of an action is irrelevant. After the action rule is created, one can add various conditions that need to hold in order for the action to be selected. For our purposes, the only relevant conditions are, the current state of the dialogue option must match the current state of the world as well as that the dialogue option must be a valid dialogue which is a FAtiMA specific rule that needs to hold for any speak actions.

After the decision-making process of a character has been formed, to ensure that the above described rules hold, additional parts of the tool need to be explored and adapted. Firstly, each character has a knowledge base where any variables and their initial values can be stored. Additionally, level of certainty can be assigned to the information stored in the knowledge base to better model real-life, however this is not currently relevant to our scenario. For our purposes, the characters need to remember the initial state of the dialogue. They could also store the emotional information relevant to the user. This was originally implemented, however Polydorou has chosen not to utilise this, and rather store the current state in their code due to easier integration. Secondly, a world model shared across the characters needs to be created. The world model in FAtiMA is responsible for storing any consequences of relevant actions. These consequences include changing values of certain variables, such as those stored in the knowledge base of a selected character. The only variable relevant to our characters is the current state of the dialogue. Therefore, the only consequence for speak action with any parameters is, changing the value of the cur-

rent dialogue state to the next state associated with the selected dialogue option, for both characters. And lastly, the actual dialogue needs to be implemented utilising the dialogue editor.

To implement a dialogue option in the editor, a dialogue action is created and added to the current dialogue actions. A dialogue action as mentioned consist of the current state, next state, meaning and style variables as well as associated text to be said. When implementing the SAT scenario, the following approach was adopted, to differentiate between dialogue actions applicable to the therapist and the player, the style variable is assigned to either therapy or play respectively. Based on the style variables, in the character decision making, we specify that the only relevant dialogue actions for that character are the ones with the applicable style. Furthermore, in terms of creating the states, the following naming conventions are followed. The first state of any of the four stages is called Stagex-Start, with x indicating the stage number from 1 to 4. The only exception is for Stage 4 which consists of multiple exercises which is named Stage4-Start, but the beginning of the exercises is referred to as Stage4x-Start, with x referring to the exercise from A to F. Afterwards, each dialogue action within a stage is numbered from 1 up, replacing the Start part in Stagex-Start. To illustrate, the first dialogue option of the first stage has the current state set to Stage1-Start and the next state set to Stage1-1. Finally, the last state that is reached with the last dialogue option is called End which is required for FAtiMA. All of the implemented dialogue actions and their assigned variables can be seen in Appendix B and the progression of states can be seen in Figure 4.1.

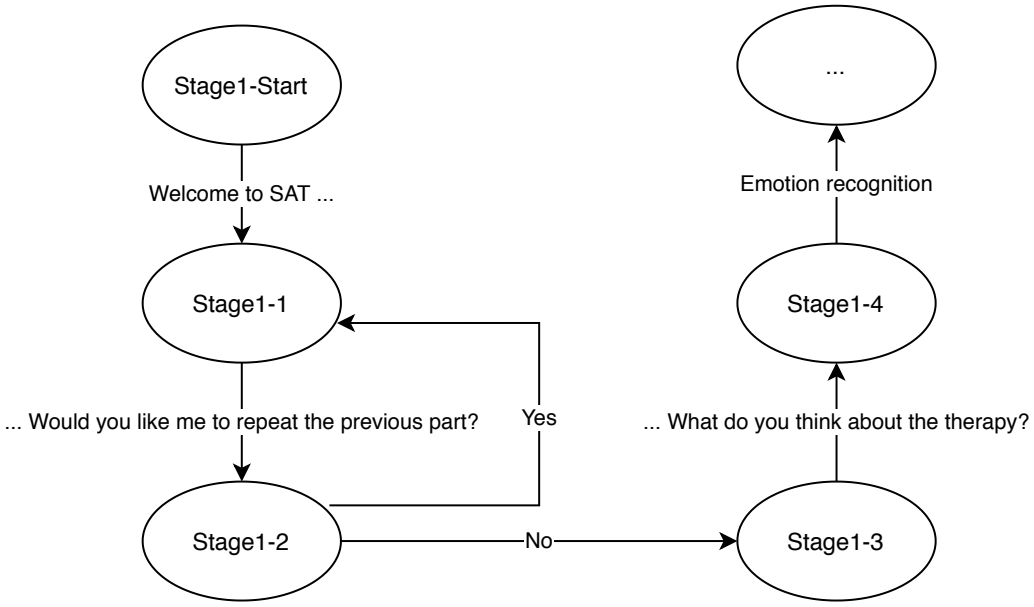

Figure 4.1: An example of state progression in Stage 1.
'...' refers to a continuation of the dialogue. Please see full dialogue in Appendix B

To make sure that this implementation is working correctly, FAtiMA tool provides these two helpful features. Firstly, a validate button within the dialogue editor, which is responsible for checking whether all states mentioned are reachable. This was quite useful to identify any possible mistypes or missing dialogue actions. Secondly, it provides a simulator which can in strides simulate the scenario. In the simulator you can choose to have agency over one of the characters, in our scenario the player, while the other character is responsible for selecting own decisions. Using the simulator, any mistypes in the dialogue options were identified as well as incorrect rules or consequences. After the completion of the scenario, the simulator was run following all paths to the end to ensure that it was correctly implemented.

4.4 C# Implementation

Even though the FAtiMA Integrated authoring tool sped up the process of implementing the scenario tremendously, we had to explore how to use the created scenario with the FAtiMA library in order to integrate it within Unity. After extensive research of the materials provided by the authors, a decision was made to create a console-based application, due to no prior experience with Unity nor access to the VR SAT implementation. The purpose of the console-based application was mainly to aid in debugging and ensuring that everything runs correctly. Once that was done, any changes agreed on with Polydorou would be made to ensure a smooth integration. The first version was a rough draft just to explore the various methods mentioned in the tutorials provided by FAtiMA authors. However, the finalised version was reformatted in line with object-oriented programming.

The implementation includes a DecisionTool class, which when constructed, loads the scenario implemented in the Integrated Authoring tool provided by FAtiMA, as well as the word model and all the characters. Additionally, the current state of the dialogue can be set to any string and therefore instead of having to run the whole scenario, one can select a starting point. Following, a scripted method is called which runs the whole scenario until the terminal state is reached. Within this loop, it is checked whether the characters have any applicable actions they can perform in that state. As within our scenario, there is always only one action per state per character, it is not possible for both characters to be able to perform an action at a same time. Whenever an applicable action is identified, a per character function is called.

If it is a player-specific function, firstly the number of dialogue options is counted, as sometimes the player instead of speaking may select from Yes and No as a response. Whenever the count is greater than 1, instead of free typing the player can select one of the options displayed to them. However, when there is only one dialogue option, the player is asked to type in their response. If the original text inputted through the FAtiMA tool corresponds to “emotion recognition”, emotion recognition is run using the player’s current input. As at the time, it was not yet known how the emotion recognition model would be integrated within this framework and the

platform, only a random number generator was used to select an emotion, for debugging purposes. Following, the world model is updated through another function, to make the consequences of the applied actions true.

If it a therapist-specific function, the applicable action gets selected. Whenever this option consists of an utterance "I see that you are", which follows after emotion recognition was triggered, the emotion detected gets added as an adjective to that utterance. Similarly, the world model updates itself based on the action. For the purposes of the console-based application both of the functions return a string describing the following, which agent (player or therapist) says what (utterance) towards target (player or therapist) in what state (current state). When the implementation was complete and ran correctly as determined through the use of the console-based application, the integration with Polydorou began.

4.5 Integration

The integration of the above implemented SAT scenario, the emotion recognition model and the overall VR SAT application was handled in close collaboration with Polydorou. The first thing to do was to decide how to integrate the emotion recognition model within the VR platform, specifically using Oculus Quest. The most straightforward way is to call the python script through the code within Unity whenever needed. This could be achieved through IronPython, which is a modification of Python for the .net framework. However, to do so Python needs to be installed on the device, which is not possible for Oculus Quest as it has an Android operating system. After this unsuccessful attempt, Polydorou came up with the idea to use a server on which the Python code would be run, and server calls would be made within their platform.

To implement this, using Flask, a micro web framework, data will be passed to the Python script, which would run per utterance emotion recognition and output the detected emotion/emotions. Making per utterance prediction is advantageous in this scenario as it allows for multiple calls to the server, which can be triggered even before the previous emotion recognition has finished. Additional advantage of utilising a server is, that less memory will be taken up by the application on the device, as the emotion recognition model, which is quite larger, will be stored on one's computer. The next step was to decide what form of data would be passed to the emotion recognition script. As the model needs both audio and transcript as input, the options are following. Either both audio file and its transcript will be given, which would mean that the transcript needs to be obtained within a C# script before engaging the server-based Python. Another possibility is, only passing the audio file and having the Python script be responsible for transcribing it. We have agreed on the latter, as it seems easier to implement automatic transcription in Python as it is likely to have more libraries capable of this due to its prevalent use in machine learning. Therefore, the Unity code would only be responsible for obtaining and passing along an audio recording of the user.

Before passing the data, a decision had to be made of how to record an utterance within Unity. The user's monologue begins after the therapist has stopped speaking and ends once the user has stopped speaking as determined by Polydorou and meanwhile, per utterance recordings are made. However, as Unity does not have a native functionality to determine when one utterance ends and another begins, Polydorou has decided to make several recordings of a certain length which would represent pseudo-utterances. We were asked to determine what this length should be and have decided on 7 seconds. The reasoning was, that to get the best results from the model, the recording should be of similar length to the average length of the utterance in the dataset. Following, several issues were identified when trying to pass the data to the model. Initially, the data was passed in the form of a JSON file, however this solution was undesirable as we did not find a way how to convert it back to audio in Python. Furthermore, we experimented with sending a byte array, as this format was easily converted back to audio. After overcoming some small issue, we successfully managed to pass the data from the platform to the emotion recognition model, using a byte array. Once the data is received and before the emotion recognition model is engaged, the data needs to be pre-processed.

In this scenario, pre-processing involves acquiring the text element from the audio by transcription. This was achieved quite easily by using a speech recognition library which is capable of transcribing text from audio data. The audio file is read and then various recognition engines such as Google Speech Recognition or IBM text to speech can be selected to acquire the text. Currently, we have opted to use the Google engine as it seems to be the most reliable for per utterance transcription. However, it is worth noting that this engine is based online, which can limit the usability. Although, since the integration between the virtual platform and the emotion recognition model is server-based, it is fair to assume that there is internet access. If no text is detected whether due to empty audio or the engine not recognising it, the script is interrupted and no emotion recognition follows.

Following, the audio needs to be loaded, then sampled to the same rate and normalised as the training data, as well as the text data needs to be tokenised and encoded. Afterwards, a model is created using the same parameters as the preferred model. Consequentially, the actual model is loaded into this structure from its pth file. Then the model is set into evaluation state to make sure that no gradients are being calculated to keep the memory usage low. Finally, the model is fed the data as input and produces the emotional output, which is then transformed using a sigmoid function. This is followed, by applying the custom thresholds to recognise which emotions were detected. The detected emotions need to be transformed to an appropriate type that can be processed by Flask. As lists do not meet this requirement, initially the emotions were passed as a string separated by a comma. However, at the moment Polydorou would like to only receive one prevalent emotion and rather utilise the multi-label classification in the future. To determine which emotion is the most dominant, we subtract the threshold from the predicted probability and then choose the largest value across all emotions. If the largest value is negative, it

means that no emotions were detected. This is then sent back to Polydorou's platform, which stores the per utterance detected emotions and then determines the most prevalent emotion after the user has finished talking. It is also worth mentioning, that a similar process was implemented using Tavernor's model, utilising their framework, which requires slightly different processing. Consequentially, Polydorou is able to choose and call either Tavernor's or the newly developed emotion recognition model, by just calling a different function. Possible advantages of this include slightly different emotions to be predicted, as Tavernor's model includes frustration and excitement, whereas our model is capable of recognising disgust and fear.

The second step to successful integration was to figure out how the FAtiMA implementation could work together with the platform. As Polydorou had already created a protocol manager, which takes care of the various stages of SAT, the avatars, any behavioural action and the environment, we had to find a way to merge this manager together with the C# script we had implemented. The first option was, to use the FAtiMA script as a protocol manager instead, however that would require lot of work and changes on Polydorou's part as their protocol manager was already quite developed. Therefore, after a careful consideration, we have agreed to rather integrate the FAtiMA script within Polydorou's protocol manager. As the expectation that the code will be used by calling only the functions needed was already considered, it was not hard to do so. Instead of running the whole scenario on a loop, as there are various environmental changes that need to made throughout, the player and therapist specific functions are called. Meaning, that instead of checking which character has something to say, Polydorou will just call the functions at the appropriate times when either character is expected to speak. Additionally, the dialogue states will be used to checkpoint the user's progress, to allow them to start from the last relevant stage. Within the dialogue between the characters, emotion recognition will be triggered by making a server call, following the same trigger as described in implementation of the FAtiMA script.

The last step was to finalise how to create an audio from the therapist utterances that could then be played to the user. To do this some sort of text-to-speech algorithm needed to be implemented. The options considered consisted of, firstly recording parts of the therapist dialogue, storing them on the device and playing them when applicable. Secondly, employing text-to-speech within Unity if a way was found. Thirdly, using another Python server-based script, only called whenever a therapist has something to say. And lastly, instead of recording real audio, we could use text-to-speech in Python to save all of the applicable dialogue options for the therapist and then storing them on Oculus Quest. In the end Polydorou has decided to handle text-to-speech within their platform.

4.6 Challenges

There were various challenges encountered when trying to implement the virtual therapist. Firstly, due to no previous experience with C#, some time had to be dedicated to understanding it, to implement the FAtiMA script. This was mitigated by previous experience with C-family programming languages such as C++ and therefore the time spent on this was reduced significantly. Secondly, due to no previous experience with Unity, it was more difficult to imagine what was needed for the integration. Thankfully, the integration was handled together with the aid of Polydorou and therefore was not affected by the lack of Unity knowledge. Perhaps the most time-consuming portion of implementing the virtual therapist was understanding and navigating the FAtiMA engine. Even though it had sufficient documentation, it is not widely employed and therefore there is not an extensive list of tutorials. However, as most of the scenario creation consists of devising logic-based rules, previous experience in logic-based learning allowed for smoother transition. Once the syntax was understood, it took very little time to implement and change the SAT dialogue.

Chapter 5

Evaluation

5.1 Emotion Recognition Model

The model's performance was evaluated based on the results achieved on the test set, which the model did not previously encounter, as all changes were made and assessed using the validation set. The evaluation process involved acquiring the predicted emotions by running the data through the model and comparing these predicted emotions with the ground truth. This was achieved by applying a sigmoid function over the predicted emotions and a custom threshold was applied over each emotion to see whether the emotion is detected. Additionally, the ground truth was in the form of multi-hot encoded labels. Based on these, various metrics were calculated to assist in assessing the model's overall and emotion-specific performance.

5.1.1 Metrics

Accuracy

Accuracy is defined as the percentage of correctly classified samples from all the samples predicted. More specifically it is calculated as (TP + TN) / (TP + TN + FP + FN), where TP refers to true positives, samples that are predicted and in truth are the positive class. TN refers to true negatives, samples that are predicted and actually belong to the negative class. FP refers to false positives, samples that are predicted as positive class but in truth are negative. FN refers to false negatives, samples that are predicted as negative but belong to positive class. When considering imbalanced data, accuracy is often unreliable as if everything is predicted as the more represented class the accuracy is still high, but the model has not learnt anything.

Weighted Accuracy

Weighted accuracy (WA) (Tong et al., 2017) transforms the accuracy metric to accommodate for an imbalance in a dataset. More specifically, it weights TP and TN based on the size of both the positive and negative classes. Weighted accuracy is

calculated as (TP * N / P + TN) / 2N, where P and N refer to all samples that belong to the positive and negative class respectively.

F1 Score

F1 score is seen as the harmonic mean of precision and recall. Precision refers to the percentage of correctly predicted positive samples across all predicted positives and is calculated as TP / (TP + FP). Recall refers to the percentage of correctly predicted positive samples to all sample that truly belong to the positive class and is calculated as TP / (TP + FN). F1 score is a more informative measure as it considers both precision and recall and is calculated 2 * precision * recall / (precision + recall).

ROC-AUC

The area under the curve of the receiver operating characteristic curve (ROC) is another metric that can provide informative evaluation of a model trained on imbalanced data. ROC curve is plotted as true positive rate against false positive rate. True positive rate refers to recall, whereas false positive rate is calculated as (1 – TN) / (TN + FP). AUC refers to the probability that a positive sample chosen at random is ranked higher than a negative sample. Due to this the ROC-AUC metrics is a reliable metric to use for imbalanced data. Additionally, when compared to F1 score which is calculated after selecting a certain threshold, meaning that the sample is classed as positive when greater than the threshold, ROC-AUC is threshold independent. More specifically, ROC-AUC is calculated for all probability thresholds. Additionally, to report an overall ROC-AUC score for a model, two methods can be considered. Macro-averaged metric does not consider the data imbalance and rather calculates the value for each class, which is then averaged. Micro-averaged metric takes into account the class imbalance by using overall TP, TN, FP, FN aggregated from all classes for calculation.

5.1.2 Results

Three final models were selected to be evaluated on the unseen test data and the results of this evaluation can be seen in Tables 5.1, 5.2 and 5.3. Considering the overall performance on all emotion classes, we can see that that the SoftW model perform slightly better than the other two. This is more pronounced when considering the F1 score, whereas overall weighted accuracy remains similar when compared to the baseline model. Accuracy is not considered in this judgement as it is biased and unreliable metric when considering imbalanced data. The SoftF model, on the other hand, seems to perform worse on average.

Model	Happiness			Sadness			Anger			Surprise		
Metrics	A	F1	WA	A	F1	WA	A	F1	WA	A	F1	WA
Baseline	68.1	73.0	67.1	**66.0**	47.4	65.0	**70.5**	49.2	**67.6**	**70.7**	26.8	64.4
SoftW	**73.7**	**80.1**	**70.0**	62.4	**52.2**	**65.3**	63.5	**52.3**	66.8	70.6	**30.5**	**65.1**
SoftF	71.9	77.9	69.5	60.7	50.0	62.8	67.4	52.0	66.8	67.6	27.4	62.1

Table 5.1: Total number of samples per emotion class.
A refers to accuracy, F1 to F1 score, and WA to weighted accuracy.

Model	Disgust			Fear			Other		
Metrics	A	F1	WA	A	F1	WA	A	F1	WA
Baseline	**82.2**	56.0	75.7	73.8	28.8	**69.4**	**58.1**	**32.6**	**61.9**
SoftW	74.8	**56.2**	**76.6**	**75.8**	**29.7**	65.4	NA	NA	NA
SoftF	73.3	52.7	73.3	60.2	21.1	50.2	NA	NA	NA

Table 5.2: Total number of samples per emotion class.
A refers to accuracy, F1 to F1 score and WA to weighted accuracy.

Model	Average			
Metrics	A	F1	WA	AUC
Baseline	**71.9**	44.8	67.3	74.1/80.8
SoftW	70.1	**50.2**	**68.2**	**74.9/82.3**
SoftF	66.9	46.8	64.1	71.5/80.5

Table 5.3: Total number of samples per emotion class.
A refers to 6-class accuracy (not including other), F1 to F1 score, WA to weighted accuracy and AUC to macro/micro ROC-AUC.

However, it is not only important to consider the overall performance of the model but also emotion specific results are informative. From this we can see that including soft labelling in the training improved F1 scores for most emotions in both Soft models. The exception would be disgust and fear in SoftF model, where the performance decreased significantly. This is especially interesting as due to the notion of focal loss, it could suggest, that in the training set some of the sample belonging to disgust and fear were easy to classify. And by placing a greater weight on the harder to classify examples, the easy ones contribute less to the loss used to update parameters in the model.

When considering adding weights to combat class imbalance the effect was marginal, as there is not significant effect observed in the smaller classes. However, the combination of < 1 weights and the soft labelling did improve the baseline model's performance by a decent amount especially when considering F1 score. The only exception would be disgust, which remained almost the same and fear where the performance increased slightly on F1 score but decreased more visibly on WA. This could imply that the average intensity for certain emotions, whose performance increased after

the use of soft labels, was higher and resulted in greater loss if misclassified than the average intensity of fear and disgust.

Overall, when considering class-specific performance, we can also observe a trend where including more class samples in the training correlates with higher performance. Such occurrence would normally be expected for the baseline model, but it should be mitigated to some extent by using the different techniques for the imbalanced data. We do not observe this in our results and therefore we may conclude that the techniques did not significantly influence the class imbalance.

Additionally, we can examine the effect of including the other/no emotion class in the baseline model. As can be seen, the overall performance decreases with all metrics when including this class. It is however understandable, that the performance on this class is worse than the other emotions, as the other/no emotion category could imply any other existing emotion. This class cannot be understood as neutral, as it is only present when none of the other basic emotions are. Additionally, this class is not present in the other two models as its inclusion significantly lowered the performance of the models on validation set. This is interesting as one would expect that including this class could still provide the model with valuable information when learning, even if it could not learn to predict that specific class well. This was not the case in our training, and it seemed that the model was rather more confused when this class was included.

To summarise, we select the SoftW model as the best performing one from the three presented with the per emotion confusion matrices displayed in Table 5.4. This choice is made specifically due to the increase of F1-scores, while still maintaining a WA similar to the of baseline model. However, if one would be interested in prioritising the performance of a certain class, such as fear one could rather use the baseline model at the expense of loss in performance on the other emotions.

Happiness			
		Actual	
		P	N
Predicted	P	2103	641
	N	402	814

Sadness			
		Actual	
		P	N
Predicted	P	814	1174
	N	315	1657

Anger			
		Actual	
		P	N
Predicted	P	794	1169
	N	277	1720

Surprise			
		Actual	
		P	N
Predicted	P	256	980
	N	185	2539

Disgust			
		Actual	
		P	N
Predicted	P	640	833
	N	165	2322

Fear			
		Actual	
		P	N
Predicted	P	202	774
	N	183	2801

Table 5.4: Per emotion confusion matrices.
P refers to the positive class and N to negative.

5.1.3 Comparison with Prior Work

As mentioned, the SoftW model was selected as the best performing, and in this section we will compare its performance with prior work on CMU-MOSEI dataset. Through this we can better evaluate the model's performance in wider context. It is worth noting that the results of the prior work were evaluated on the same test set as suggested by the authors. The comparison of our performance to the prior work can be seen in Tables 5.5 and 5.6.

Model	Happiness			Sadness			Anger			Surprise		
Metrics	A	F1	WA	A	F1	WA	A	F1	WA	A	F1	WA
Zadeh	-	66.3	66.3	-	66.9	60.4	-	72.8	62.6	-	85.5	53.7
Akhtar	-	67.0	53.6	-	72.4	61.4	-	**75.9**	66.8	-	**86.0**	60.6
Chauhan	-	71.3	51.9	-	**72.6**	61.8	-	74.7	**67.4**	-	**86.0**	58.2
Beard	-	-	-	-	-	-	-	-	-	-	-	-
Delbrouck	67.1	78.1	-	**82.7**	31.4	-	**81.7**	28.4	-	**90.5**	15.8	-
SoftW	**73.7**	**80.1**	**70.0**	62.4	52.2	**65.3**	63.5	52.3	66.8	70.6	30.5	**65.1**

Table 5.5: Comparison of SoftW model to prior work.
A refers to accuracy, F1 to F1 score, and WA to weighted accuracy.

Model	Disgust			Fear			Average		
Metrics	A	F1	WA	A	F1	WA	A	F1	WA
Zadeh et al. (2018)	-	76.6	69.1	-	**89.9**	62.0	-	76.3	62.3
Akhtar et al. (2019)	-	**81.9**	72.7	-	87.9	62.2	-	78.6	62.8
Chauhan et al. (2019)	-	81.8	74.1	-	87.8	63.9	-	**79.0**	62.9
Beard et al. (2018)	-	-	-	-	-	-	-	-	61.6
Delbrouck et al. (2020)	**79.1**	25.5	-	**88.2**	26.7	-	-	-	-
SoftW	74.8	56.2	**76.6**	75.8	29.7	**65.4**	**70.1**	50.2	**68.2**
T-test	-	-	-	-	-	-	-	-	**0.0076**

Table 5.6: Comparison of SoftW model to prior work.
A refers to accuracy, F1 to F1 score, and WA to weighted accuracy. T-test with $p < 0.05$ shows that the results are statistically significant compared to best performing model (Chauhan et al., 2019).

In terms of overall performance of the SoftW model, the model performs significantly better in terms of WA as determined by a t-test with regards to the best performing prior work (Chauhan et al., 2019) with $p < 0.01$. However, our model suffers in performance when considering F1 score. This is quite interesting, especially considering the better performance on WA and the difference between our average F1 and those of the other models. Both metrics are important to consider when evaluating a model and depending on the focus one can prefer one or the other. More specifically, F1 only takes into account the values of TP, FP and FN, whereas weighted accuracy takes into account TN as well. As none of the mentioned papers provide their confusion matrices it is hard to understand how they managed to obtain such high F1 scores while only mediocre WA. In our scenario, the low F1 scores are associated mostly with the underrepresented classes as to detect majority of these, the model classifies a great amount of FP as well.

When considering class-specific performance, the results of our model in happiness detection is significantly higher when considering both F1 and WA metrics. Therefore, based on the presented prior work, our happiness detection comprises the state-of-the-art. However, when considering the other emotions, on one hand none of the associated F1 scores manage to compete with the best performing models. On the

other hand, consulting WA our model outperforms the other models on almost all emotions, with the exception of anger which is marginally worse. Moreover, we can compare the accuracy of our model with the one work (Delbrouck et al., 2020) that provides this metric, where except for happiness our model performs worse. Although, it should be noted that this specific prior work also reports very low F1 scores, which are significantly lower than ours. This likely means, that their model is classifying majority of samples as the more prevalent class, which in this dataset is the negative class for each emotion, except for happiness. Since we were selecting thresholds to maximise F1 and WA on our validation set, our accuracy is not as high. This could be easily adapted by changing the thresholds and achieve very high accuracies in a similar manner to that prior work. This however does not imply that the model performs well due to the imbalance in the dataset and is an undesirable behaviour.

To summarize, our model significantly outperforms the mentioned state-of-the-art models on happiness recognition and suffers on the other emotions in terms of F1 scores, due to high number of FP. However, as it performs significantly better than the other models regarding WA, we can consider its performance as competitive. It is also worth noting that the best performing prior works all utilise three modalities, including video as well. Therefore, we can expect the performance of our model to increase if we were to include this additional modality in the future.

5.1.4 Limitations

Despite the achieved performance, there are various limitations to our model and if these were to be addressed in the future, it is likely that the performance would increase further.

Firstly, the high rate of false positives, as seen in the confusion matrices should be mentioned. In order to classify a majority of the positive samples correctly the threshold had to often be lowered, which resulted in additional negative samples to be classified as positive. This is unwanted behaviour as we would like to avoid predicting one's emotional state incorrectly. Due to the relation of the low F1 scores to the imbalance in the dataset, we could expect an increase in performance, if the imbalance was addressed further. Even if certain techniques to combat this were attempted, it appears that they did not had the effect we expected. In terms of class weights there does not necessarily seem to be anything to improve on and we can conclude that this technique in our scenario was not very impactful. However, experimenting with the focal loss should be explored in the future. More specifically, tuning the alpha and gamma values further could potentially increase the performance. As well as one could try to include the class-specific loss weights, as possibly a combination of these could have an effect. Additionally, the focal loss trick as mentioned by Mao (2020) could be worth implementing. This trick comprises of adapting last layer bias term initialisation so that the less represented samples are initially incorrectly classified. Such adaption would result in much higher loss values

for such samples and therefore the network would learn better and faster the less represented classes. Whereas, normally it takes time to see the effect of focal loss, and in the meantime the model can begin overfitting some of the more represented or easier to classify examples. An additional technique that should be attempted even if it likely will not have a large impact, is oversampling, which would result in a more balanced dataset. However, the weighting of losses should have similar impact to oversampling and therefore we do not expect a significant change when attempting this.

Secondly, as our model begins to overfit the training data after only a short number of epochs, one should attempt to find the likely cause. Within our model, this could potentially be the BERT architecture used to process the text data. Due to the large number of parameters the model may quickly learn the necessary relations between text and emotions, whereas the rest of our model may not be as quick when considering the audio data. This could result in not maximising the performance of our model as perhaps the audio part has not been trained as well as BERT. This could be mitigated in the future by considering separate parameters for the two sub-models such as including a smaller learning rate for BERT in comparison to the audio model. Additionally, one could experiment with dropout probabilities for the BERT model to prevent overfitting.

Thirdly, the hyperparameters could be tuned more extensively, by experimenting with a more diverse range of values. As easy as this would be in terms of implementation, it would also be very time consuming due to the model's training time. However, if one would have access to multiple GPU's, the training could potentially be sped up and therefore make it easier to tune said hyperparameters.

Lastly, if one was interested in a different application of this model, where all three modalities, including video could be utilised, it would be worth to include this modality in the training. Especially, as including all three modalities seems to outperform audio and text in CMU-MOSEI prior work. As the authors of the CMU-MOSEI dataset provide already pre-extracted features for the video, it would not be too hard to adapt the model to process such data. However, if one was interested in using the raw audio data, a deep residual network could be used, followed by LSTM as in Tzirakis et al. (2017).

5.2 SAT Scenario Implementation

One of the aims of this project was to implement a certain middle layer that would integrate the emotion recognition model together with the interaction between the user and the virtual therapist. To evaluate how well this was achieved, one should first consider the various aspects of this implementation, more specifically the dialogue manager and the said middle layer.

In terms of actually implementing the scenario in the form of a dialogue, as mentioned this was conducted using the Authoring Integrated tool provided by FAtiMA. Consequentially, the full scenario dialogue was implemented which is advantageous. If only a part of the scenario was implemented, this would limit the potential use of the overall application and would have to be improved upon in the future. The choice of using the FAtiMA toolkit is beneficial when considering future extensions of the scenario. As SAT is relatively novel, the scenario is constantly evolving and therefore it is reasonable to expect that the implemented dialogue may need to be adapted in the future. Therefore, when selecting the dialogue manager engine, it was important that it easily allows such changes. FAtiMA provides the Integrated Authoring tool which means that the scenario can be changed without accessing any code, which allows for someone non-proficient in programming to adapt the dialogue. As well as to simplify this process further, one can import the dialogue from an Excel file, including the various variables needed for a dialogue option. Consequentially, even the interaction with the FAtiMA tool can be limited to certain extent. Overall, when considering the dialogue management, the FAtiMA toolkit was well selected as it allowed to implement the whole scripted scenario in a desirable order, as well as the scenario can be easily adapted in the future.

5.2.1 Limitations

When evaluating how well the scenario was implemented one should also consider the associated disadvantages and any potential improvements.

Firstly, when considering the scenario itself, the responses of the virtual therapist are purely scripted. Therefore, if one uses the application on multiple occasions, as it should be used in order to see the potential benefits, they may notice that the therapist has no agency in terms of their responses. On one hand, implementing the scenario purely scripted allows to be exactly aware of the responses of the virtual therapist, which may be desirable when considering the interaction between the therapist and vulnerable population. On the other hand, it can limit the real-world experience that we are trying to emulate within the VR administered SAT. Perhaps, the user may feel less connected to the environment and therefore not fully commit to SAT, which may affect its efficacy. One possible way to address this is to include ML in the process of forming the dialogue, by having the therapist generate their responses rather than to consider only the scripted options. However, this is a potential risk, as we will not be exactly aware of what the therapist may say. If the therapist were to say something insensitive, this may not influence the healthy population, but it could potentially have negative impact on people with mental disorders. Another option that could be considered, is to include multiple dialogue options per current state, with the therapist being able to choose from the scripted responses. FAtiMA allows for such case and even includes the possibility of either random choice or a choice which is determined based on various variables, while still including the randomised aspect. As a result, the therapist would respond differently on multiple occasions during the same portion of the dialogue, which would create an illusion

of agency, while still relying on purely scripted answers.

Secondly, FAtiMA does not allow for variable dialogue options, more specifically to have an utterance of the form of a set sentence with inserting a variable option within this utterance. This is specifically important to SAT scenario whenever the therapist asks for confirmation of the detected emotion from the user, as it is of the form "I see that you are [emotion], is that right?". If we would consider only utilising the FAtiMA tool in implementing such utterance, either the utterance would have to be split or multiple dialogue options would have to be created accounting for every potential detected emotion. It is possible, but it greatly increases the number of dialogue options as well as current states. Additionally, it would mean that if a different emotion recognition model was added, capable of recognising other emotions, additional dialogue options would have to be added. To combat this, together with Polydorou we have decided to only include the utterance "I see that you are" and add the rest of the utterance within our code, dependent on the detected emotion. However, this is perhaps not the cleanest solution as when looking only on the implemented scenario in the tool, one would not understand why a part of the utterance is missing. Perhaps, in the final version of the application, it would be more beneficial to add the various dialogue options, if one is aware of all the emotions that could be detected.

Thirdly, the overall potential of the FAtiMA toolkit is not fully utilised. We are only using it as a dialogue manager, whereas it comprises of modules which could enhance the experience even further. However, as the SAT scenario is very straightforward with every dialogue option following the other, with certain exceptions such as repeating the tutorial etc., the variables stored in the world model only include the current state of the world. Whereas other variables, such as the emotion of the user, are stored within the platform rather than the world model. This is certainly fulfilling its function for now, however if one would be interested in making the application even more interactive, the other functionalities should be explored. One possibility is to include the child avatar as a character and to form their decision making to respond to the various decision and behaviours of the user. In addition, one could store the emotion state of the user and based on the current emotion or even the user's progress, the responses of the therapist could vary. Moreover, as one of the modules is autobiographic memory, certain behaviours of the user could have consequences in the future. Consequentially, all of the behaviours of the user, therapist and the child could be modelled through FAtiMA rather than the platform and interact with the platform by the changes within the environment only.

5.3 Integration

Due to the cooperation with Polydorou and the object-oriented formatting of the C# implementation, the integration with Polydorou's platform progressed quite smoothly and quickly. However, some of the small issues could have been avoided if we were to discuss the integration sooner and format our contributions in line with that. Overall, we have opted to not use certain FAtiMA functionalities for easier integration, whereas these could have allowed for a more efficient result. However, we managed to merge the functionalities of Polydorou's platform together with the C# implementation successfully, resulting in a functional VR administered SAT application.

When considering the integration of the emotion recognition model, one should evaluate the choice of using a server to run the model. An advantage of this method is the easily achievable per utterance prediction, which can be run simultaneously as the user is speaking, saving the computing power of the virtual reality device. This is especially beneficial due to the actual time it takes to predict a 7 seconds long utterance, which is the utterance length that gets passed to the model. On average, it takes approximately 12 seconds to run the server code to detect emotions, whether Tavernor's or the newly trained model is used. This was observed using a personal computer with Intel i7-8565U CPU. As the server is engaged throughout the user's monologue, which is likely to consist of several utterances, we can expect that by the time the user finishes speaking, most of the emotion recognition on the previous utterances has finished. And therefore, one does not have to wait to acquire the prediction of the last utterance, as already some data on the user's emotional state has been gathered.

Another advantage is that the user does not have to store all the necessary files needed to run emotion recognition on the Oculus Quest device. As the device is untethered, it features storage of either 64 or 128 GB, which is smaller than that of most personal computers. The size of the files used for the emotion recognition is approximately 7 GB, which could get larger if we would decide to incorporate more models, as one model is approximately 1.5 GB in size. Such amount of data could be considered large to store on the Quest devices, especially the ones with 64GB memory. If a server is used to run emotion recognition, whether hosted on a personal computer or hosted on cloud (e.g. AWS), it is likely that the size of the data will be less of a problem. Furthermore, hosting the emotion recognition on a server is independent of the virtual reality device. Therefore, if in the future one would like to utilise the emotion recognition in other devices, it would be possible without any changes to server code.

Furthermore, considering the implementation of the server-based emotion detection, one is able to utilise two different models. More specifically, the one trained by Tavernor on IEMOCAP and the one trained on CMU-MOSEI. Instead of having to access the code, the model-specific emotion recognition can be run just by accessing a different Flask route. Not only this allows for different emotion sets predictions

based on the model, but the two models could also be run simultaneously, and both of their outputs could be potentially considered. Moreover, it is quite easy to add additional models trained using or based on Tavernor's framework if desirable.

5.3.1 Limitations

As this is not necessarily the final product when considering the VR administered SAT application, it is important to evaluate its limitations and how these could be improved.

Perhaps the biggest drawback is the time it takes to run the server-based emotion recognition script. As mentioned, this is mitigated if the user's response consists of several utterances, however one should consider the possibility of user only speaking for a short time. More specifically, if the user would just talk for 7 seconds or less, which is the average utterance length, they would have to wait for the emotion recognition to finish. Therefore, there would be a small period of time, where both the user and the therapist would be silent, and the user would wait for the therapist to say something. This could potentially have negative impact on the user's experience, although to conclude this, one would have to conduct an experiment to acquire the opinions of the potential users. When considering how this could be handled in the future, there are several options.

On one hand, another emotion recognition model could be trained which would be smaller and faster, such as instead of using BERT-Large, as suggested by Tavernor (2020) one could rather utilise a smaller BERT architecture. However, one should note that this option is time consuming and would also likely result in a less accurate model. On the other hand, one could try to mitigate this within the platform. To appropriately assess one's emotion state, using one utterance may not give reliable results. Therefore, the user could be motivated by the therapist to continue speaking, if they decide to only speak for short amount of time. Another option could involve the therapist saying something, meanwhile the rest of the emotion recognition finishes. Consequentially, even if the user would still have to wait for few seconds at least they would know that the therapist has acknowledged their input. It is also possible that if a more powerful computing device were to be used to host the server, this would result in faster emotion recognition. However, further testing would need to be done to confirm this.

Another limitation of the server-based approach is that currently we rely on the user to store all the necessarily files for emotion recognition and host the server on their personal computer. This is specifically problematic as the server would have to be set up each time prior to the application being run. Therefore, for the final product one should consider hosting a cloud-based server, which would handle requests from the application from all the users. However, if this method was to be implemented, one should consider encrypting the data passed to the server so that the privacy of the users was protected, in line with the applicable data protection laws.

Furthermore, in this current version the server passes along only the most dominant emotion to the platform. This is a significant disadvantage when using the multilabel model as trained on CMU-MOSEI, as the data for the other predicted emotions does not get utilised. As in real life scenario a person often feels multiple emotions, not utilising this data reduces the model's external validity to some extent. Therefore, in the future implementations, one should consider adapting the code to pass back multiple detected emotions and perhaps their associated probabilities or intensities. This would allow the platform to make a more accurate judgement when considering emotion recognition per monologue.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this project, we provide two important components and assist in integrating them within the developed VR administered SAT application.

Firstly, an adapted emotion recognition model based on Tavernor's (2020) work was trained to be capable of multi-emotion recognition, utilising the CMU-MOSEI dataset. We can expect the model to have higher external validity, due to its multi-label nature, with the model being able to detect multiple universal emotions. Additionally, the detection of all the basic emotions; happiness, sadness, anger, surprise, disgust and fear, is desirable as it encompasses the majority of the emotional states that the user is likely to experience. More specifically, despite the low prevalence of fear in many datasets, our model manages to detect this emotion with a sufficient accuracy. Consequently, this could be especially helpful when treating anxiety through SAT, as the model could detect higher levels of fear in the user. Such data could be used to assess the impact of SAT on the disorder as well as potentially further personalise the application, if in the future disorder specific protocols are developed. Another factor that should increase the model's external validity is the choice of dataset. As real-life data was used to train the model, we can expect it to generalise better to new unseen data, than a model trained on an acted dataset. This is an important advantage, as the model's intended use is to detect emotions of the users of the VR SAT application. Regarding the model's performance, it achieves significantly better results on happiness recognition than prior work, and competitive results while recognising the rest of the basic emotions. When considering the design choices together with the achieved performance, the produced model is very suitable for assisting the VR SAT application.

Secondly, a dialogue manager has been developed, which purpose is to handle the dialogue aspect of SAT scenario, to allow the virtual therapist to efficiently guide the user through the therapy. The inclusion of a virtual human providing support and responses to the user, should allow for more innate interaction within the application and therefore may have positive impact on the effects of the therapy. Due to the

scripted nature of the therapist responses, the user should receive proper guidance to fully experience the therapy. At the same time, these responses only fulfil their intended purpose without judging or insensitively responding to the user's actions. The management of the therapist responses and the overall dialogue was implemented using the FAtiMA toolkit. The choice of utilising such toolkit is very advantageous, as whenever the scenario needs to be updated, it can be done by simply adapting an excel file.

To conclude, due to the successful integration of the above described components within the virtual platform developed by Polydorou (2020), the virtual therapist willl be able to guide the user through the therapy as well as appraise user's emotional state when needed by interacting with the emotion recognition model. As a consequence, the detected emotions can now be displayed on the child avatar, which should positively affect the user's ability to bond and interact with the child and commit to the therapy. Overall, the inclusion of this work within the final product is likely to benefit the user and therefore amplify the potential of SAT in VR.

6.2 Future Work

When devising the final version of the VR SAT application, several adaptations to this work should be considered to maximise its impact.

Firstly, one should assess the results of the per utterance emotion recognition using a different dataset. More specifically, collecting real-life data resembling the actual input that the model will receive when deployed, should allow to evaluate the performance in real life setting. During this project, an effort was made to acquire such data, however due to the associated timeframe, it is currently a work in progress. To fairly judge the model, such data should include participants responses to similar questions, as provided in the scenario prior to emotion recognition. Such evaluation will likely be informative and reveal whether the external validity is indeed as high as expected.

Secondly, the model should be evaluated in terms of per utterance prediction time on different devices to assess whether its speed can be maximised. If the end result is still too slow for real-time emotion recognition, some of the mitigation strategies as outlined in evaluation should be considered. Perhaps, instead of retraining the model with different architecture, which would be time consuming, one could instead focus on maximising the user's monologue, as more data would lead to a more reliable prediction. However, if one chooses to retrain the model, further strategies should be considered to improve its performance.

As the model performs very well on happiness, which is the most prevalent emotion, we could expect an increase in performance when supplying the model with more data on the less represented emotions. As CMU-MOSEI is the largest dataset up to date for emotion recognition, instead of selecting a new dataset, one could combine

multiple datasets for training. This would require them to be processed in a similar manner which will be time consuming. However, as this model does not utilise extracted features, but uses raw data, all datasets with raw audio and text could be considered, and only the emotional labelling would have to be merged or adapted. Before attempting this, one could also focus on hyperparameter tuning and experimenting with the different techniques that can be used on imbalanced datasets, as outlined in evaluation of the model.

Lastly, one could explore the functionalities of FAtiMA to greater extent in order to improve the user's experience further. More specifically, the whole protocol could be handled within the toolkit rather than using custom protocol manager and a dialogue manager. This would be beneficial, as it would make for a more complete integration, but it would require a further understanding of the tool. As mentioned, one could in addition to the user and therapist, model the child and its reactions to provide further interactions. Perhaps the child could also speak, dependent on their emotional state or simply react differently to the user, not by just visibly feeling the emotion.

Bibliography

Ainsworth, M.D.S, Blehar, M.C., Waters, E., & Wall, S. (1978). Patterns of attachment: A psychological study of the strange situation. Hillsdale, NJ: Erlbaum.

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. arXiv preprint arXiv:1905.05812.

Al-Omari, H., Abdullah, M. A., & Shaikh, S. (2020, April). EmoDet2: Emotion Detection in English Textual Dialogue using BERT and BiLSTM Models. In 2020 11th International Conference on Information and Communication Systems (ICICS) (pp. 226-232). IEEE.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: Author.

Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-emotion neural network for emotion recognition from text. IEEE Access, 7, 111866-111878.

Beard, R., Das, R., Ng, R. W., Gopalakrishnan, P. K., Eerens, L., Swietojanski, P., & Miksik, O. (2018, October). Multi-modal sequence fusion via recursive attention for emotion recognition. In Proceedings of the 22nd Conference on Computational Natural Language Learning (pp. 251-259).

Bowlby, J. (1969). Attachment and Loss: Vol.1. Attachment. London: Hogarth Press Bretherton, I., & Munholland, K. A. (2008). Internal working models in attachment relationships: Elaborating a central construct in attachment theory.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335.

Candilis, P. J., & Pollack, M. H. (1997). The hidden costs of untreated anxiety disorders. Harvard Review of Psychiatry, 5(1), 40-42.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE

transactions on affective computing, 5(4), 377-390.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019, November). Contextaware interactive attention for multi-modal sentiment and emotion analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5651- 5661).

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019, November). Contextaware interactive attention for multi-modal sentiment and emotion analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5651- 5661).

Cittern, D., Edalat, A., & Ghaznavi, I. (2017). An immersive virtual reality mobile platform for self-attachment. AISB.

Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. Emotion, 18(1), 75. Davies, R. C., Johansson, G., Boschian, K., Lindén, A., Minör, U., & Sonesson, B. (1998). A practical example using virtual reality in the assessment of brain injury. In 2nd European Conference on Disability, Virtual Reality and Associated Technologies, Skovde, Sweden.

Delbrouck, J. B., Tits, N., Brousmiche, M., & Dupont, S. (2020). A Transformer-based jointencoding for Emotion Recognition and Sentiment Analysis. arXiv preprint arXiv:2006.15955.

Demyttenaere, K., Bruffaerts, R., Posada-Villa, J., Gasquet, I., Kovess, V., Lepine, J., ... & Kikkawa, T. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. Jama, 291(21), 2581-2590.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Difede, J., & Hoffman, H. G. (2002). Virtual reality exposure therapy for World Trade Center post-traumatic stress disorder: A case report. Cyberpsychology & behavior, 5(6), 529-535.

Edalat, A. (2015, July). Introduction to self-attachment and its neural basis. In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Edalat, A. (2017). Self-attachment: A holistic approach to Computational Psychiatry. In Computational neurology and psychiatry (pp. 273-314). Springer, Cham.

Ekman, P. (1992). Are there basic emotions?.

Emmelkamp, P. M., Bruynzeel, M., Drost, L., & van der Mast, C. A. G. (2001). Virtual reality treatment in acrophobia: a comparison with exposure in vivo. CyberPsychology & Behavior, 4(3), 335-339.

Eng, W., Heimberg, R. G., Hart, T. A., Schneier, F. R., & Liebowitz, M. R. (2001). Attachment in individuals with social anxiety disorder: the relationship among adult attachment styles, social anxiety, and depression. Emotion, 1(4), 365.

Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Svetlana Kiritchenko and Saif M. Mohammad. In Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM), New Orleans, LA, USA, June 2018.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92, 60-68.

Fodor, L. A., Coteț, C. D., Cuijpers, P., Szamoskozi, Ș., David, D., & Cristea, I. A. (2018).
The effectiveness of virtual reality based interventions for symptoms of anxiety and depression: A meta-analysis. Scientific reports, 8(1), 1-13.

for the Revision, I. A. G. (2011). A conceptual framework for the revision of the ICD-10 classification of mental and behavioural disorders. World Psychiatry, 10(2), 86.

Gartlehner, G., Wagner, G., Matyas, N., Titscher, V., Greimel, J., Lux, L., ... & Lohr, K. N. (2017). Pharmacological and non-pharmacological treatments for major depressive disorder: review of systematic reviews. BMJ open, 7(6), e014912.

Ghaznavi, I., Jehanzeb, U., Edalat, A., & Gillies, D. (2019). Usability evaluation of an immersive virtual reality platform for self-attachment psychotherapy.

Gorman, J. M. (1996). Comorbid depression and anxiety spectrum disorders. Depression and anxiety, 4(4), 160-168.

Gratch, J., Hartholt, A., Dehghani, M., & Marsella, S. (2013). Virtual humans: a new toolkit for cognitive science research. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770- 778).

Heaney, D. (2019, June 28). Facebook's Prototype VR Face Tracking Got Even Better. Retrieved from https://uploadvr.com/frl-multiview-face-tracking/.

Honyashiki, M., Furukawa, T. A., Noma, H., Tanaka, S., Chen, P., Ichikawa, K., ... & Caldwell, D. M. (2014). Specificity of CBT for depression: a contribution from multiple treatments meta-analyses. Cognitive therapy and research, 38(3), 249-260.

Huang, C., Trabelsi, A., & Zaïane, O. R. (2019). ANA at SemEval-2019 Task 3: Contextual Emotion detection in Conversations through hierarchical LSTMs and BERT. arXiv preprint arXiv:1904.00132.

Joselson, N., & Hallén, R. (2019). Emotion Classification with Natural Language Processing (Comparing BERT and Bi-Directional LSTM models for use with Twitter conversations).

Kant, N., Puri, R., Yakovenko, N., & Catanzaro, B. (2018). Practical Text Classification With Large Pre-Trained Language Models. arXiv preprint arXiv:1812.01207.

Klinger, R. (2018, August). An analysis of annotated corpora for emotion classification in text. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2104-2119).

Laura Ana Maria Bostan and Roman Klinger. A survey on annotated data sets for emotion classification in text. In Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics, Santa Fe, USA, August 2018.

Laynard, R., Clark, D., Knapp, M., & Mayraz, G. (2007). Cost-benefit analysis of psychological therapy. National Institute Economic Review, 202(1), 90-98.

Li, Y., Zhao, T., & Kawahara, T. (2019, September). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Interspeech (pp. 2803-2807).

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980- 2988).

Liu, P., Du, C., Zhao, S., & Zhu, C. (2019). Emotion Action Detection and Emotion Inference: the Task and Dataset. arXiv preprint arXiv:1903.06901.

Livingstone, S. R., & Russo, F. A. (2018). 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.

Main, M., & Solomon, J. (1990). Procedures for identifying infants as disorga-

nized/disoriented during the Ainsworth Strange Situation. Attachment in the preschool years: Theory, research, and intervention, 1, 121-160.

Mao, L. (2020). Use Focal Loss To Train Model Using Imbalanced Dataset. Retrieved from https://leimao.github.io/blog/Focal-Loss-Explained/.

Mascarenhas, S., Guimarães, M., Prada, R., Dias, J., Santos, P. A., Star, K., ... & Kommeren, R. (2018, August). A virtual agent toolkit for serious games developers. In 2018 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-7). IEEE.

Mehrabian, A., & Russell, J. A. (1974). An approach to environmental psychology. the MIT Press.

Mikulincer, M., & Shaver, P. R. (2019). Attachment orientations and emotion regulation. Current Opinion in Psychology, 25, 6-10.

Moors, A. (2009). Theories of emotion causation: A review. Cognition and emotion, 23(4), 625-662.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H., & Dokania, P. K. (2020). On using focal loss for neural network calibration.

Murphy, B., & Bates, G. W. (1997). Adult attachment style and vulnerability to depression. Personality and Individual differences, 22(6), 835-844.

Norton, P. J., & Price, E. C. (2007). A meta-analytic review of adult cognitivebehavioral treatment outcome across the anxiety disorders. The Journal of nervous and mental disease, 195(6), 521-531.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Pfefferbaum, B., & North, C. S. (2020). Mental health and the Covid-19 pandemic. New England Journal of Medicine.

Phillips, I. (2014). Lack of imagination: Individual differences in mental imagery and the significance of consciousness. In New waves in philosophy of mind (pp. 278-300). Palgrave Macmillan, London.

Polydorou, N. (2020). User-friendly highly interactive virtual reality platform for selfattachment therapy (Unpublished master's dissertation). Imperial College London, London, United Kingdom.

Qin, A. (2018). A Pytorch implementation of Focal Loss. Retrieved from https://www.kaggle.com/c/salt-identification-challenge/discussion/65938.

Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013, April). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG) (pp. 1-8). IEEE.

Rizos, G., & Schuller, B. W. (2020, June). Average Jane, Where Art Thou?–Recent Avenues in Efficient Machine Learning Under Subjectivity Uncertainty. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 42-55). Springer, Cham.

Rizos, G., Baird, A., Elliott, M., & Schuller, B. (2020, May). Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3502-3506). IEEE.

Rizzo, A. A., Buckwalter, J. G., Neumann, U., Kesselman, C., & Thiébaux, M. (1998). Basic issues in the application of virtual reality for the assessment and rehabilitation of cognitive impairments and functional disabilities. CyberPsychology & Behavior, 1(1), 59-78.

Romano, D. M. (2005). Virtual reality therapy. Developmental medicine and child neurology, 47(9), 580-580.

Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.

Sharac, J., Mccrone, P., Clement, S., & Thornicroft, G. (2010). The economic impact of mental health stigma and discrimination: a systematic review. Epidemiology and Psychiatric Sciences, 19(3), 223-232.

Stein, D. J. (2013). What is a mental disorder? A perspective from cognitive-affective science. The Canadian Journal of Psychiatry, 58(12), 656-662.

Stein, D. J., Phillips, K. A., Bolton, D., Fulford, K. W. M., Sadler, J. Z., & Kendler, K. S. (2010).

What is a mental/psychiatric disorder? From DSM-IV to DSM-V. Psychological medicine, 40(11), 1759-1765.

Stolzenburg, S., Freitag, S., Evans-Lacko, S., Speerforck, S., Schmidt, S., & Schomerus, G. (2019). Individuals with currently untreated mental illness: causal beliefs and readiness to seek help. Epidemiology and psychiatric sciences, 28(4), 446-457.

Tavernor, J. (2020). Cross-corpus Speech and Textual Emotion Learning for Psychotherapy (Unpublished master's dissertation). Imperial College London, London, United Kingdom.

Tong, E., Zadeh, A., Jones, C., & Morency, L. P. (2017). Combating human trafficking with deep multimodal models. arXiv preprint arXiv:1705.02735.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5200-5204). IEEE.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301-1309.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

World Health Organization. (1992). The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines. Geneva: World Health Organization.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). 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) (pp. 2236-2246).

Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. (2018, February). Multiattention recurrent network for human communication comprehension. In Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence (Vol. 2018, p. 5642). NIH Public Access.

Appendices

SAT Scenario (Polydorou, 2020)

Stage I: Introduction to Self-attachment therapy

Virtual agent: Welcome to the Self-Attachment therapy. My name is Ana and I will be your assistant throughout the therapy. Firstly, you will have to familiarize yourself with the platform.

[Here we will have a tutorial for the user.]

Virtual agent: You can repeat the tutorial as many times as you wish as soon as we finish with the introduction. Now that you have familiarized yourself with the platform, I will give you a brief introduction to the Self-Attachment therapy.

Virtual agent: The Self-Attachment therapy is based on the Attachment Theory from John Bowlby which suggests that from the first year of life, children create an emotionally attachment with their primary care-giver. The type of attachment determines the personality and emotional development of an individual in their adult life, as well as the way they perceive the world. Different types of attachment objects are utilised by a securely attached individual in order to feel secure in a stressful situation. Self-Attachment therapy is employed in order to help you feel the love you were deprived of in your childhood, and create a secure attachment that will allow you to control your emotions and cope with the distressing situation. In this therapy you are imagined to consist of an childhood self, representing your emotional self, which becomes dominant under stress, and an inner adult corresponding to your logical self, which is dominant in the absence of stress. The aim of the therapy is the creation of an affectional bond between the childhood self and the adult self who takes the role of a new primary carer-giver. In this way, a secure attachment will be created between you and the child inside the platform. You will, also, be guided through all the stages of the Self-Attachment therapy.

Virtual agent: Would you like me to repeat the previous part?

User: [YES/NO]

Virtual agent: Okay, let us move on.

Virtual agent: What do you think about the therapy?

User: [Here we give the user the chance to talk so we can do emotion recognition.]

Virtual agent: It is important to believe in this therapy and be committed. Are you sure you can do that?

User: [YES/NO]

Virtual agent: Very well.

Virtual agent: [Terms and conditions.] Do you accept the terms and conditions?

User: [YES/NO]

Platform design

Initially, the user is found in an empty room with the virtual agent. This room will be in an old house that cannot provide safety and represents the inner world of the individual who suffers from a mental disorder. By the end of the therapy this old house must be replaced by a robust house which is a safe haven representing an individual with secure attachment. The building process will reflect the progress of the patient who is trying to repair their damaged inner world and create a better life.

In this stage the introduction and tutorial take place. The aim of the tutorial is to allow the user to learn the navigation controls and how to interact with the platform. Also a video that shows how someone can do self-massage is playing on the TV.

Stage II: Connecting with the childhood self

Virtual agent: Now I will ask you to enter the child's room, where you have to conceptualise the childhood self, to develop empathy and then compassion towards it. Initially, the child will be happy, thus by looking at it try to recall happy moments from your childhood. After you do that you can turn off the light with the light switch next to the door. When you turn off the light the child will become sad. Therefore, try to recall sad memories from your childhood. You can turn the light on/off as many time as you want. Try to connect with the child. Take your time. I will be here.

[At any time the user can exit the room and talk to the virtual coach again.]

Virtual agent: Hello again. Did you manage to connect with the child?

User: [YES/NO]

Virtual agent: How are you feeling?

User: [User talking. Emotion recognition.]

Virtual agent: I see that you are [Detected emotion]. Is that right?

User: [YES/NO]

Platform design

The user will be in the same old room as in Stage I. In the next room will be the child avatar and on the wall will be two images taken from the individual's childhood. One photo will be a happy memory of the individual and the other photo will be an unhappy memory. Initially, the room will be dark and the unhappy photo will be visible next to the sad child. Inside the room there is a light switch which when the user turns on the happy photo will replace the unhappy one and the child will become happy. After the individual recalls their happy and unhappy memories, the child will leave the room and enter a brighter kids' room. By following the child, the user shows a connection with it and a bond is initiated.

Stage III: Falling in love with the childhood self

Virtual agent: Now, I would like you to enter the room again and sing your favorite love song to the child in order to make it happy. This will help you to create an affectional bond and fall in love with the child. You can go whenever you are ready.

User: [The user enters the room and sings. We can probably predict arousal and valence level from the singing. User exits the room.]

Virtual agent: How are you feeling now?

User: [User talking. Emotion recognition.]

Virtual agent: Now you have to adopt the child and vow that you will support and protect it. Are you sure you can do that?

User: [YES/NO]

Virtual agent: You are ready for the main part of the therapy. You will have to interact with the child as a good parent in order to minimise the negative emotions and maximise the positive affects. This can take multiple sessions, which allows you to leave and enter the platform as many times as you want. I will keep track of your progress.

Platform design

The user enters the new child's room and loudly sings a song. This will change the emotional state of the child from neutral to happy and then to dancing. If the user stops singing, then the child will get sad so as to show to the user that must keep singing. After the user exits the room, they will notice that the walls of their new house were built and thus the development of the new house has begun.

Stage IV: Developmental exercises for the childhood self

Virtual agent: Good morning/afternoon, how are you today?

User: [User talking. Emotion recognition.]

Virtual agent: I see that you are [Detected emotion]. Is that right?

User: [YES/NO]

Virtual agent: Let's start with the different stages of the therapy.

Every time the user enters the platform, they will be in this main stage. The sky will be dark at the start and after the first three stages it will become sunny with mountain view.

Type A: Sessions for processing the painful past

Virtual agent: Try to recall a traumatic episode in your childhood with as much details as possible. Try to remember a different event from previous sessions. How did you feel? Did you feel fear, helplessness, humiliation or rage?

User: [User talking. Emotion recognition. The predicted emotion is projected to the child avatar.]

Virtual agent: When you are ready, please enter the room and parent your childhood self. You can loudly reassure and embrace the child to show support as a good parent. For example, say with a loud voice, "Why are you hitting my child?" and "My darling, I will not let them hurt you any more." Also, you can cuddle your childhood self by giving yourself a massage. All these will help to change the emotional state of the child and thus your emotional state.

User: [Enters the room. We can do emotion recognition on the reassuring talk of the user. At any time the user can exit the room.]

Initially, the child will have negative emotions (anger, sad, fear), depending on the predicted emotional state of the user. Then, after being reassured, the child will have neutral emotions and after being embraced the child will be happy.

Type B: Sessions to process the current negative emotions

Virtual agent: Now, what are your most recent negative emotions? Are they related to family, friends, work, education or social affairs? Try to feel these emotions right now.

User: [User talking. Emotion recognition.] Virtual agent: How are you feeling? Do you feel anger, rage or fear?

User: [User talking. Confirmation of predicted emotions. The predicted emotion is projected to the child avatar.]

Virtual agent: Now enter the room and parent your childhood self. You can loudly reassure and embrace the child to show support as a good parent. Also, you can cuddle your childhood self by giving yourself a massage. All these will help to change the emotional state of the child and thus your emotional state.

User: [Enters the room. We can do emotion recognition on the reassuring talk of the user. At any time the user can exit the room.]

Initially, the child will experience negative emotions (anger, sad, fear) depending on the predicted emotional state of the user. Similar to the Type A stage, the emotional state of the child will change to neutral and then to happy.

Type C: Protocols for creating zest for life

Virtual agent: Now enter the room and sing your favourite love song to the child. Also, try shaking your head and shoulders and moving your eyes, eyebrows, hands and arms.

User: [Enters the room. Spends some time interacting with the child and exits the room.]

Virtual agent: Singing will help you contain your negative affects and experience real joy. Thus, you have to repeat these exercises under many different circumstances, such as when you are walking at the streets or when you are working.

At this stage, the child's emotion will be neutral, however, after the song the child will be happy and then will start dancing. The favorite song of the user will be playing in the background.

Type D: Getting over the negative emotions

Virtual agent: On the wall you can see an image of the Gestalt vase. This vase represents the negative emotions and the more you look at it, the more you get drowned into its negativity.

User: [Looking at the image.]

Virtual agent: However, after successful completion of the previous exercises you have created a strong pattern of love and when you look at the image again, you

will discover two white faces that represent your childhood self and adult self who face each other.

User: [Looking at the same image.]

The Gestalt vase will be on the wall with white background so that the vase is more obvious. Then, the white background will be reduced so that the two faces will be more obvious to the user.

Type E: Socialising protocols for the childhood self

Virtual agent: Giving yourself a massage will help you improve more your emotional state.

User: [Self-massage.]

Virtual agent: Do you feel better now?

User: [YES/NO]

Virtual agent: You have successfully competed all the exercises for this session. It is very important now to apply the things you have learned in your real-life through interactions with other people. Can you do that?

User: [YES/NO]

Type F: Creating a more optimal internal working model

Virtual agent: At the end of each session you will notice that a piece of your new beautiful house has been added. After many sessions you will have a complete house which will be your secure haven. This house represents your new optimal internal working model for interpreting and managing your relationships with others. Your developed internal working model will help you have peace with yourself and with other people.

Virtual agent: You have reached the end of this session. How are you felling now?
User: [User talking. Emotion recognition will show us the improvement of the emotional state of the user.]

Virtual agent: Very well. See you next time. Goodbye.

After each session a piece of the house will be built, starting with the walls and then adding furniture piece by piece, and finally creating a yard full of trees.

FAtiMA implementation of the dialogue

An example of the formatting needed to import a dialogue option into FAtiMA Integrated Authoring Tool can be seen in Table 1.

Current State	Next State	Meaning	Style	Utterance
Stage2-3	Stage2-4	-	therapy	How are you feeling?

Table 1: An example of a dialog option

To see all the dialogue options and their assigned variables implemented in FAtiMA, please refer to the SAT.xlsx file submitted with the software/code archive.

Instructions on how to run the server-based emotion recognition

1. Go to the Integration folder in the submitted software/code archive.
2. Install all dependencies specified in the provided README file.
3. Set up flask by running the commands:
   (a) set FLASK_APP=server.py
   (b) set FLASK_ENV=development
4. Run flask with the command:
   (a) flask run
5. Copy the link displayed in console by flask and open it in your browser.
6. To test the server on the sample data select either multilabel or multiclass classification by adding '/multilabel' or '/multiclass' to the previously copied address.
7. The server-based script is called within SAT VR application created by Neophytos Polydorou. If you would like to run this application, please request access at neophytos.polydorou19@imperial.ac.uk.

Ethics Checklist

	Yes	No
Section 1: HUMAN EMBRYOS/FOETUSES		
Does your project involve Human Embryonic Stem Cells?		x
Does your project involve the use of human embryos?		x
Does your project involve the use of human foetal tissues / cells?		x
Section 2: HUMANS		
Does your project involve human participants?		x
Section 3: HUMAN CELLS / TISSUES		
Does your project involve human cells or tissues? (Other than from "Human Embryos/Foetuses" i.e. Section 1)?		x
Section 4: PROTECTION OF PERSONAL DATA		
Does your project involve personal data collection and/or processing?		x
Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)?		x
Does it involve processing of genetic information?		x
Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc.		x
Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets?		x
Section 5: ANIMALS		
Does your project involve animals?		x
Section 6: DEVELOPING COUNTRIES		
Does your project involve developing countries?		x
If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned?		x
Could the situation in the country put the individuals taking part in the project at risk?		x
Section 7: ENVIRONMENTAL PROTECTION AND SAFETY		
Does your project involve the use of elements that may cause harm to the environment, animals or plants?		x
Does your project deal with endangered fauna and/or flora /protected areas?		x
Does your project involve the use of elements that may cause harm to humans, including project staff?		x
Does your project involve other harmful materials or equipment, e.g. high-powered laser systems?		x
Section 8: DUAL USE		
Does your project have the potential for military applications?		x
Does your project have an exclusive civilian application focus?		x
Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items?		x

Does your project affect current standards in military ethics – e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons?		x
Section 9: MISUSE		
Does your project have the potential for malevolent/criminal/terrorist abuse?		x
Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery?		x
Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied?		x
Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project?		x
SECTION 10: LEGAL ISSUES		
Will your project use or produce software for which there are copyright licensing implications?		x
Will your project use or produce goods or information for which there are data protection, or other legal implications?		x
SECTION 11: OTHER ETHICS ISSUES		
Are there any other ethics issues that should be taken into consideration?	x	