

Imperial College London

MENG INDIVIDUAL PROJECT

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Towards an Optimal Model of Empathy and Emotion Recognition for Self-Administrable Self-Attachment Techniques

Author:
Radostin Petrov

Supervisor:
Dr. Anandha Gopalan

Second Marker:
Prof. Abbas Edalat

August 3, 2025

Abstract

A rising trend in mental health illness and distress worldwide has increased the demand for digitalisation of mental health services. A novel theory in psychotherapy called self-attachment theory is focused on dealing with emotional distress by reflecting on childhood relationships with primary-caregivers. This protocol-based therapy makes it adaptable to a virtual environment showing promise in improving patients' health.

This work improves the existing framework of a self-attachment therapy chatbot by increasing the emotional and empathetic capabilities of the chatbot. It involves collecting a novel corpus of utterances that contains 3 times more emotions than in the previous version. Inspired by research into existing work, the dataset is used in training state-of-the-art language models, that are applied to emotion and empathy classification tasks. These models are combined with the existing platform to create a new version of a self-attachment chatbot that shows pronounced performance in recognising emotion and reacting with empathy.

The results of this chatbot framework are evaluated in a non-clinical trial, where participants are guided through self-attachment protocols in their interaction with the chatbot. The outcomes show a positive receipt of the framework and favourable improvements in the main two goals of this work - increasing the emotional and empathetic capabilities of the chatbot.

This work concludes with an overview of the conducted research and contributions and assesses its strengths and shortcomings when compared to the previous versions.

Acknowledgements

First and foremost, I would like to thank Dr. Anandha Gopalan for supervising this project and guiding me through the obstacles confronted in a final year project. I express my gratitude for the patience and understanding provided and a never-ending encouragement that helped see this work to its end.

I would also like to thank Prof. Abbas Edalat, whose passion for the research of self-attachment therapy has been greatly motivational for me and has given this project a deeper meaning for me. His assistance in providing resources and most importantly the ability to fund a crowd-sourcing has taught me a lot about academic research, which is going to be a one of the biggest takeaways from this degree for me.

I extend my deep gratitude to Lisa Alazraki for the endless support in this project. She was always happy to offer her help and her work was a building stone for the entire research and it would not be an exaggeration to say that without her contributions this project would have not taken shape.

In addition, I would like to acknowledge everyone in the Algorithmic Human Development group at Imperial College London, who dedicate their work in this area of combining novel psychotherapy methods with developments in the area of computing and natural language processing.

Lastly, I would like to give my deepest gratitude to my family. Their support has been crucial for my progress not only in this project but the whole degree. I would like to thank my parents for believing in me, when it was difficult for myself to do so. Needless to say, that without them this work would not have been possible as I would have not attended a university as distinguished as Imperial. The challenges of taking the bold step of studying here has shaped me into the person I am and I would gladly say that I have no regrets for undertaking this life step.

This project has been an insightful experience and I am glad it is what I will be sending off my degree with.

Contents

1	Introduction	7
1.1	Motivation	7
1.2	Project Aims	8
1.3	Outcomes and Contributions	8
2	Background	10
2.1	Preliminaries	10
2.1.1	Emotion Theory	10
2.1.2	Empathy	11
2.1.3	Natural Language Processing	11
2.2	Literature Review	14
2.3	Existing Framework	16
2.4	Ethical Considerations	17
3	Data Collection	18
3.1	Process of Data Collection	18
3.2	Survey Creation Process	19
3.2.1	Respondent Recruitment	19
3.3	Data Analysis	20
3.3.1	Demographic Distribution of the Data	20
3.3.2	Empathetic Utterances	21
4	Implementation	22
4.1	Data Augmentation	22
4.1.1	Downsampling	22
4.1.2	Back Translation	22
4.1.3	Wordnet Synonym Word Replacement	23
4.1.4	Data Augmentation Comparison	23
4.2	Emotion Classification	24
4.2.1	Fine-Tuning	25
4.2.2	Model Evaluation	26
4.3	Empathy Classification	26
4.3.1	Cross-model Semi-supervised Learning Configuration	26
4.3.2	Mutli-objective Optimisation Formula	29
4.4	Hyperparameter Tuning	29
5	Non-clinical trial	31
5.1	Trial Setup and Limitations	31
5.2	Trial Results	32

6	Project Evaluation	34
7	Conclusion and Future Work	38
A	Self-Attachment Technique Protocols	46
B	Analysis of Emotion Datasets in Research Literature	52
C	Conversation Flow	56
D	Example Survey Entries from Data Crowd-Sourcing	58

List of Figures

2.1	Plutchik’s proposed three-dimensional emotional circumplex, describing the relationship between complex emotions[16]	11
2.2	Architecture of a transformer model, as proposed by Vaswani et al. [35]	13
2.3	Examples of two dialogues with a respective emotional context from the EMPATHETICDIALOGUES dataset[20].	16
2.4	User interface of the self-attachment technique chatbot platform [15].	16
3.1	An example survey focused on the loving emotional context.	19
3.2	Data distribution of age groups	21
3.3	Data distribution of gender	21
4.1	Teacher-Student Pipeline	28
5.1	Distribution of trial responses across engagement, fluency and overall usefulness.	32
5.2	Distribution of responses on the chatbot’s emotion recognition.	33
5.3	Distribution of responses on the chatbot’s empathy.	33
5.4	Distribution of emotions expressed by participants in the trial.	33
6.1	Comparison of trial responses rating the chatbot’s emotion recognition.	34
6.2	Comparison of trial responses rating the chatbot’s empathy.	34
6.3	Comparison of trial responses assessing the platform’s engagement.	35
6.4	Comparison of trial responses assessing the platform’s usefulness.	36
B.1	Distribution of emotions in the GOEMOTIONS dataset based on number of examples and showing the interrater correlation by annotators [19].	52
B.2	Results from the experiments on fine-tuning a pre-trained BERT model with GOEMOTIONS [19].	53
B.3	Distribution of emotions in the EMPATHETICDIALOGUES dataset based on their representation and showing the most frequently appearing speaker and listener words [20].	54
B.4	Evaluation results from several configurations run with the EMPATHETICDIALOGUES dataset based on a 4-layer custom transformer architecture [20]. Metrics include precision retrieving the correct test candidate out of 100 test candidates (P@1,100), average of BLEU-1,-2,-3,-4 (AVG BLEU) and perplexity (PPL).	55
C.1	Graph of the conversation flow in a positive emotional context [15].	56
C.2	Graph of the conversation flow in a negative emotional context [15].	57

D.1	Example of an accepted and quality response from the crowd-sourcing of disgust and disappointment.	58
D.2	Example of low quality response from the crowd-sourcing of disgust and disappointment. This entry was accepted based on the rejection criteria but the prompts for the emotion were deleted and considered as empty as they did not represent the emotional context necessary. .	58
D.3	Example of a rejected low quality response from the crowd-sourcing of disgust and disappointment.	59

List of Tables

3.1	Distribution of collected samples for each of the 8 new emotions. *Note that the conversation flow for positive emotions such as "love" is considerably shorter and the number of rewriting required are lower.	20
4.1	Comparison of performance between different configurations of ROBERTA _{BASE} fine-tuned on an augmented corpus. All scores are in percentages of a single unit.	24
4.2	Comparison of performance between different configurations of ROBERTA _{BASE} fine-tuned on an augmented corpus. All scores are in percentages of a single unit.	24
4.3	Comparison of performance between different configurations on the emotion classification task. All scores are in percentages of a single unit. *Note that given the unavailability of different jealousy and envy labels in EMPATHETICDIALOGUES, the used emotions from the EMPATHETICPERSONAS12 corpus are now 11.	26
4.4	Comparison of performance between models trained on true-labelled data and semi-supervised trained ones. All scores are in percentages of a single unit.	29
4.5	Parameters selected for tuning and the best performing combination for the classification tasks.	30
6.1	Comparison of the macro-averaged F1 scores for the best models in emotion and empathy classification between the previous and the new version of the chatbot. All scores are in percentages of a single unit. .	37

Chapter 1

Introduction

1.1 Motivation

Mental health is an important, yet complicated issue in current times. Worrying trends show an increase in mental health illness and diagnosis worldwide which has considerable impact on all areas of life [1]. Recent data also exhibits disturbing signs of lasting negative psychological impact that the COVID-19 pandemic has had on children with a rapid increase in post-traumatic stress disorder diagnosis in the past few years [2, 3]. Developments in psychotherapy and psychology suggest that negative early childhood attachment can leave children susceptible to mental health illness due their impressionability [4, 5].

Traditional mental health services are projected to struggle in meeting the growing demand in the coming years[6]. One of the biggest challenges with current methods is that the understanding of a mental illness is specific to the individual. Particularly these services fail to provide adequate help to patients in low- and middle-income families [6, 7]. This creates an opportunity for the development of accessible automated therapy methods.

Virtual psychotherapy allows an affordable alternative that can help a person suffering from mental health [8]. Conversational and relational agents have already been deployed in the field of psychotherapy with the aim to simulate face-to-face therapy sessions [9]. The current psychological methods that are employed in these systems follow the cognitive-behaviour and psychodynamic approaches in psychotherapy. These approaches reduce emotional distress by targeting disorder-specific symptoms and applying cognitive interventions [10]. However, these forms of therapy lack focus on a patient’s past experiences, bonds and events, formed in the earlier childhood years, which is central to a paradigm in psychotherapy known as self-attachment theory [11]. Self-attachment is a form of psychotherapy that focuses on the relationship with care-givers during infancy of a human. It has proven successful in identifying and battling mental illness in pre-clinical trials and has been suggested as a viable digital option for virtual psychotherapy [12, 13].

The current work in automating the self-attachment technique has contributed to the creation of a virtual platform for interacting with a chatbot therapist [14]. The platform has been curated for self-attachment therapy by guiding a patient through the self-attachment protocols based on their emotional distress. The chatbot has been improved in terms of understanding emotional contexts and observing

empathy using classification techniques from modern research in the field [15]. The existing framework has been received positively by users taking part in non-clinical trials[15], however, in its current version, the chatbot recognises a limited amount of emotions from the ones targeted by the self-attachment technique [12]. Enhancing the the emotional and empathetic capabilities of the current models is suggested as an improvement that could increase the engagement and usefulness of the chatbot and contribute towards an optimal virtual psychotherapist.

1.2 Project Aims

The work aims to improve the current framework of the virtual self-attachment technique chatbot [15] by studying developments in natural language processing in the tasks of emotion and empathy classification. More specifically, the objectives of this project are to increase the number of emotions that can be recognised by the chatbot and to achieve an optimally empathetic and non-offensive agent for self-attachment therapy. This is accomplished by crowd-sourcing domain-specific data to create a novel emotional corpus that applies to the context of self-attachment therapy and its protocols. By training state-of-the-art language models on the collected dataset, the performance of the emotion recognition and empathy capabilities of the chatbot is improved. These improvements are evaluated in a non-clinical trial, where volunteers interact with the new version of the chatbot.

The project is split in three stages: collecting an emotion corpus, experimenting with and implementing different state-of-the-art models, and evaluating the results based on the non-clinical trial. Referring to published research in the field, a subset of 8 new emotions of relevancy to self-attachment therapy are chosen for crowd-sourcing the corpus. This process involved recruiting research workers to provide examples of different emotions by answering specially designed surveys. The resulting dataset is augmented and used with the language models in two downstream tasks - emotion and empathy classification. The performance of the models is evaluated by comparing objective metrics when being tested on a held out dataset.

The models are also compared against existing work and the best scoring ones are implemented into the chatbot framework. A non-clinical trial with 16 participants is conducted to collect feedback assessing the emotional and empathetic capabilities of the chatbot. Additionally, volunteers are asked to evaluate the platform’s usefulness and rate the engagement and fluency of the chatbot.

1.3 Outcomes and Contributions

The contributions of this work list a successfully crowd-sourced emotion and empathetic corpus, called EMPATHETICPERSONAS12, consisting of 12 labelled emotion classes. This dataset is a result of the previously collected corpus EMPATHETICPERSONAS for four emotional contexts (anger, anxiety/fear, sadness, joy) [15], combined with eight new emotions - love, insecurity, disgust, disappointment, shame, guilt, envy, jealousy. This shows a notable increase of 200% in the number of emotions adapted to the chatbot framework.

This work also provides improvements over the training methods by

implementing larger language models trained on the EMPATHETICPERSONAS12 corpus. Notable metrics show the best transformer configuration achieving 92.02% accuracy and 91.79% macro-averaged F1 score on the test set for the emotion classification task. By leveraging semi-supervised learning methods used in research, the empathy classification of the chatbot is also improved. Using a cross-model student-teacher learning strategy the chatbot is trained on 4 times as many empathetic datapoints from EMPATHETICPERSONAS12 than previously. Macro-averaged F1 score of 81.36% shows an improvement over the empathetic classification achieved in the previous version of the chatbot [15].

Finally, the overall result of the project is evaluated in the non-clinical trial collecting effective feedback across many categories. This includes better ratings in the emotional recognition and empathetic tasks of this project, as well as positive results in the overall usefulness of the platform and the engagement with the chatbot. The assessment of this work shows its beneficial contribution to the improvement of virtual self-attachment therapy and gives insight into its future developments.

Chapter 2

Background

This section aims to explore novel methods in natural language processing for emotion and empathy classification in a wide range of emotional contexts. It builds an understanding of the background subjects on emotion and empathy, specifically when applied to the topic of self-attachment therapy, as well as the preliminaries of natural language processing. Research in the area of language models is referred to during the implementation of this work and its evaluation.

2.1 Preliminaries

2.1.1 Emotion Theory

Emotion classification is a comprehensive task that contains an understanding of the basic emotion theory. Emotions play a significant part in human lives and emotional distress is a driving factor for people to pursuing psychotherapy [16]. In repairing emotional disorders, a psychotherapist must understand the roots of emotions and their significance to the human evolution.

In the field of neurobiology, studies claim that the origin of emotions and their manifestation in human lives can be explained by the role they play in survival mechanics [16]. The identification of emotions from a biological standpoint has defined several "basic" emotions that are observed in all humans and higher animals. Ekman [17] describes six basic emotional states (anger, fear, sadness, enjoyment, disgust, surprise) that differ in many dimensions with regards to antecedent events and the behavioural response in humans. Plutchik's proposition of basic emotions consists of 8 emotions formed into four bipolar pairs: happiness versus sadness, anger versus fear, acceptance versus disgust, surprise versus expectancy [16]. Additionally, he proposes a three dimensional model, called a circumplex, that extends over 32 separate emotional states (Figure 2.1). This representation shows the colour-coded groups of basic emotions with certain emotions outside of the coloured spectrum, emphasising that certain "complex" emotions are a combination of basic ones [16].

Both Ekman's and Plutchik's works have been influential in existing research into emotion classification and the construction of emotion corpora, targeted at generating human-like empathetic discourse [18, 19, 20].

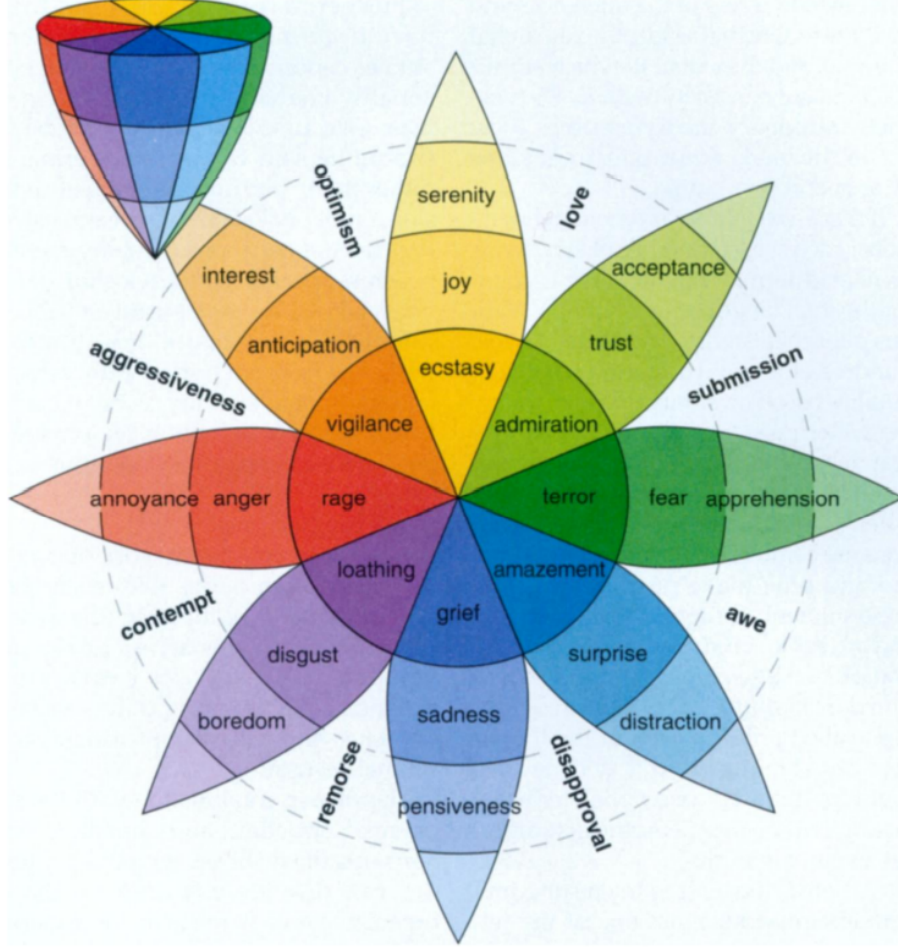


Figure 2.1: Plutchik’s proposed three-dimensional emotional circumplex, describing the relationship between complex emotions[16]

2.1.2 Empathy

The connection between emotion and empathy is established in the works of Carl Rogers, who defines empathy as the process of understanding the experiences of others and the capacity to respond emotionally [21, 22]. The role of empathy in psychotherapy is found to be the most significant aspect of the productive relationship between a patient and a therapist [23]. Engaging and reacting with empathy to emotional conversations have proven to beneficially increase the outcomes in human-to-human interaction [24, 25, 26]. The effect of empathy in the context of natural language tasks has shown that empathetic conversational agents produce higher satisfaction ratings and trustworthiness in users across multiple human-computer interaction domains including virtual psychotherapy [27, 28, 29, 30].

2.1.3 Natural Language Processing

Natural language processing (NLP) is a field in computer science and more specifically machine learning that studies the processing of language by computers for a specific task. Having its foundation in automata theory [31] and information theory [32] developed in the 1950s, NLP has seen a steady improvement in the

recent years moving from rule-based linguistic models in the 1960s-1970s to deeper analysis with corpus-based methods in the 1980s-1990s [33]. Since the turn of the century, the linguistic representations of NLP models have been improved with the use of statistical modelling and the rise of neural networks has contributed to rapid developments in the power of natural language models.

Language Models

Language models are statistical techniques in NLP that are used for predicting the probability of a sequence of words appearing in a sentence given the frequency of its occurrence in a given corpus [34]. For a given sequence of m words w_1, w_2, \dots, w_m , a language model computes the probability of observing the sequence $P(w_1, w_2, \dots, w_m)$ by applying the chain rule on the joint probabilities of each word in the sequence appearing, given the words before it:

$$P(w_1, w_2, \dots, w_m) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_2, w_1) \times \dots \times P(w_m|w_{m-1}, \dots, w_2, w_1)$$

Evaluating the performance of the above language model requires measuring the perplexity when applied to an unseen corpus. Perplexity is the inverse probability of the language model, so a better performing model is the one with the minimum perplexity:

$$PP(w_1, w_2, \dots, w_m) = PP(w_1, w_2, \dots, w_m)^{-1/m} = \sqrt[m]{1/P(w_1, w_2, \dots, w_m)}$$

Language models are at the basis of all NLP processes and they can be n-gram based, neural-based, or in a transformer architecture.

Transformers

Transformers are a model architecture based on a self-attention mechanism that relates different positions of a sequence to compute a representation of the sequence [35]. This form of a sequence-to-sequence architecture has a characteristic encoder-decoder structure [35] - the encoder takes a sequence of symbols (x_1, x_2, \dots, x_n) and maps it to a representation $\mathbf{z} = (z_1, z_2, \dots, z_n)$ and decoder, which generates an output sequence (y_1, y_2, \dots, y_m) of symbols [35]. Each step of the model is recurrent, consuming the previously generated symbols as new input (Figure 2.2).

Parallelising the self-attention fully connected layers of the transformer architecture allow such language models to be more efficient in training on large datasets than neural-based models, such as recurrent neural networks (RNNs) [36], which makes transformers superior in many complex NLP tasks [37]. Transformers are at the basis of modern research in emotion and empathy classification and are proven successful in dialogue generation tasks, showing engagement and achieving state-of-the-art results [19, 20, 38]. Different transformer language models have been developed, improving on the initially proposed architecture by Vaswani et al. [35]. Notable transformers considered for this project are described in the following sections and details about their resulting performance are outlined in [Chapter 4 Implementation](#).

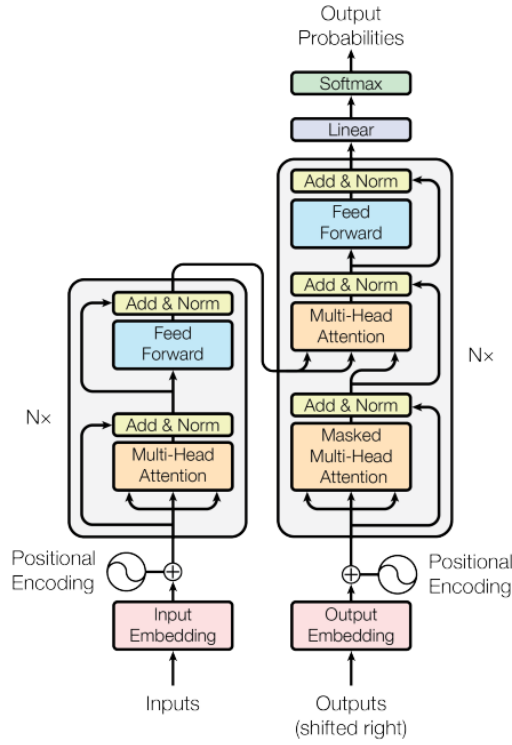


Figure 2.2: Architecture of a transformer model, as proposed by Vaswani et al. [35]

GPT-2

GPT-2 is a language model transformer trained on a 40GB diverse custom web-scraped dataset in an unsupervised manner [39, 40]. The model's main objective is the prediction of the next word in the sequence, given all preceding words. It has proven to be performant on a large amount of tasks regardless of the domain, due to its diverse corpus data [40]. It utilises a left-context decoder-only architecture [39] in contrast to the transformer model proposed by Vaswani et al.[35] and each GPT-2 output is one token that is added to the sequence of inputs, which simulates RNN architectures [36]. The input representation uses a unique BPE (Byte Pair Encoding) [41] approach, which allows for the encoding of any string, making GPT-2 applicable to any general dataset [39]. GPT-2 remains the chosen transformer model for evaluating the perplexity of a sentence's fluency used in the virtual chatbot framework, supported by research into similar tasks [42]. Originally GPT-2 consists of 1.5 billion parameters [43], however, for this work a smaller model is being used, consisting of 117 million parameters.

BERT

BERT is an encoder-only language model with a focus on achieving bidirectional pre-training language representations through utilising MLM ("Masked Language Models"). These models use word-masking as a strategy that randomly masks input tokens and predicts them based on context which creates a unique left-to-right input representation [44]. BERT and BERT-based models are designed to be fine-tuned with a single additional output layer, proving successful results in a number of downstream tasks that match state-of-the-art models [44].

RoBERTa

ROBERTA is a transformer model that improves BERT on several metrics. It leverages the left-to-right encoder-only architecture of BERT, while increasing training batches and uses a dynamic word-masking method. ROBERTA masks training inputs in different ways every four iterations during the preprocessing stage as opposed to BERT’s static masking [45]. ROBERTA also consists of 123 million parameters in its base version and 355 million parameters in its large version and outperforms BERT in most downstream tasks including multi-class classification [46, 47]. Both ROBERTA_{BASE} and ROBERTA_{LARGE} have been considered for this work and evaluated on their performance in classifying emotional data, as well as scoring empathetic context.

T5

T5 is a transformer-based language model, which follows the original encoder-decoder architecture, proposed by Vaswani et al. [35] and leverages an approach known as text-to-text processing [48]. Specifically, T5 takes text as input and returns text as output, making it adaptable to virtually any task, as opposed to BERT-based models, whose outputs limit them to classification or masking problems [48]. T5 also achieves state-of-the-art results in many tasks, including classification problems and its base model, which consists of 220 million parameters was chosen for experimentation, comparing its results to BERT-based models.

Megatron-BERT

Megatron-LM is a novel work in training transformer-based models with billions of parameters by an leveraging efficient parallel training approach [49]. This enables large language models to be accessible without domain-specific libraries or configurations [49]. The method was developed by NVIDIA and has been incorporated with state-of-the-art transformer models [50, 49] and Megatron has shown improvements over BERT-based models by producing comparably better performance with parameter sizes of 3.9 billion. Megatron is also the basis and centre of collaboration in successfully training Megatron-Turing NLG - the largest Natural Language Generation model with 530 billion parameters [51].

Megatron-LM is at the basis of developing MEGATRON-BERT - the largest BERT-based model adopting the BERT architecture. It has an increased number of layers producing a transformer model, comprising of 3.9 billion parameters [50]. In this work MEGATRON-BERT is applied for emotion classification and based on its performance is compared against the above-mentioned BERT-based models and T5.

2.2 Literature Review

There are multiple approaches to the emotion classification problem in the modern research via various language models. Statistic-based approaches such as Wang, et al.’s [52] Twitter data analysis on self-labeled tweets makes use of the WordNet Affect lexicon to extract features. A total of 32 features were crafted for each

tweet, representing 32 subcategories of positive, negative, neutral and ambiguous emotion [52]. The study concludes that a combination of n-gram language models, together with existing sentiment and emotion lexicons perform the best in emotion recognition tasks, however, lexicons are less effective in identifying complex emotions compared to semantic analysis methods [52]. Hand-crafted features are constrained by the dependency on manually created linguistic resources and require domain expertise [18]. In contrast to the aforementioned supervised methods of emotion classification, Saravia et al. [18] propose a framework of a multi-layer CNN architecture with enriched patterns. Word embeddings have been shown to perform well on classification tasks due to the preservation of semantic relationships between the patterns [18]. While the model performs well on a wide range of context, it is limited to only 8 emotions. As a result, it can be concluded that the importance of contextual meaning and semantic analysis is essential when processing text-based datasets in NLP models, regardless of dimensionality.

The availability of labelled emotional data beyond the scope of the Ekman’s basic emotions [16] has been limited due to the challenges of collecting and annotating such a corpus. Developments in this direction have been made in the form of fine-grained datasets, such as GOEMOTIONS, the largest manually labelled dataset, consisting of 58000 Reddit comments [19]. The corpus is collected following an emotion allocation, supported by psychology literature and maximises on the coverage of emotional ranges available from Reddit’s subforums [19]. The dataset contains a total of 27 emotions and includes multi-label annotations for utterances with similar emotions. The distribution of the classes in the GOEMOTIONS corpus is outlined in the Appendix.

In social media content, such as publicly available tweets [52] or Reddit threads [19], the "discussion" is conducted in front of an immeasurably wide audience. This can lead to subject matter biased towards self-presentation [20, 53] as opposed to through private channels [54], such as the ones conducted in psychotherapy. Hence, maximising performance for a context such as self-attachment technique can only be achieved through a domain-specific dataset. Rashkin et al. [20] present a domain-specific dataset with the goal of training a model responsive to any emotion. The crowd-sourced dataset consists of near 25000 conversations¹, balanced around 32 emotion labels. It consists of different situations between a speaker (presenting the situation with a specific emotion) and a listener (recognising the emotional range of the speaker and replying with empathy). Figure 2.3 shows an example dialogue for different emotional contexts. These discussions are comparably similar to the interaction between a psychotherapist and a patient and EMPATHETICDIALOGUES was used for the emotion classification task of this project. The distribution and analysis of all 32 considered emotions can be found in the Appendix.

¹The dataset is open-source and can be found at <https://github.com/facebookresearch/EmpatheticDialogues>

<p>Label: Afraid Situation: Speaker felt this when... "I've been hearing noises around the house at night" Conversation: Speaker: I've been hearing some strange noises around the house at night. Listener: oh no! That's scary! What do you think it is? Speaker: I don't know, that's what's making me anxious. Listener: I'm sorry to hear that. I wish I could help you figure it out</p>	<p>Label: Proud Situation: Speaker felt this when... "I finally got that promotion at work! I have tried so hard for so long to get it!" Conversation: Speaker: I finally got promoted today at work! Listener: Congrats! That's great! Speaker: Thank you! I've been trying to get it for a while now! Listener: That is quite an accomplishment and you should be proud!</p>
--	--

Figure 2.3: Examples of two dialogues with a respective emotional context from the EMPATHETICDIALOGUES dataset[20].

2.3 Existing Framework

The current version of the self-attachment technique chatbot ², features a virtual agent platform that specialises in recommending users self-attachment protocols throughout the interaction[15] [14]. The chatbot includes several interactive personas and utilises fine-tuned models that perform the emotion classification and empathetic responses. This version of the platform is able to recognise and react to four emotional states - joy, anger, sadness, anxiety/fear and is trained on EMPATHETICPERSONAS, a crowd-sourced corpus specialised in the domain of self-attachment therapy with regards to these four emotions [15].

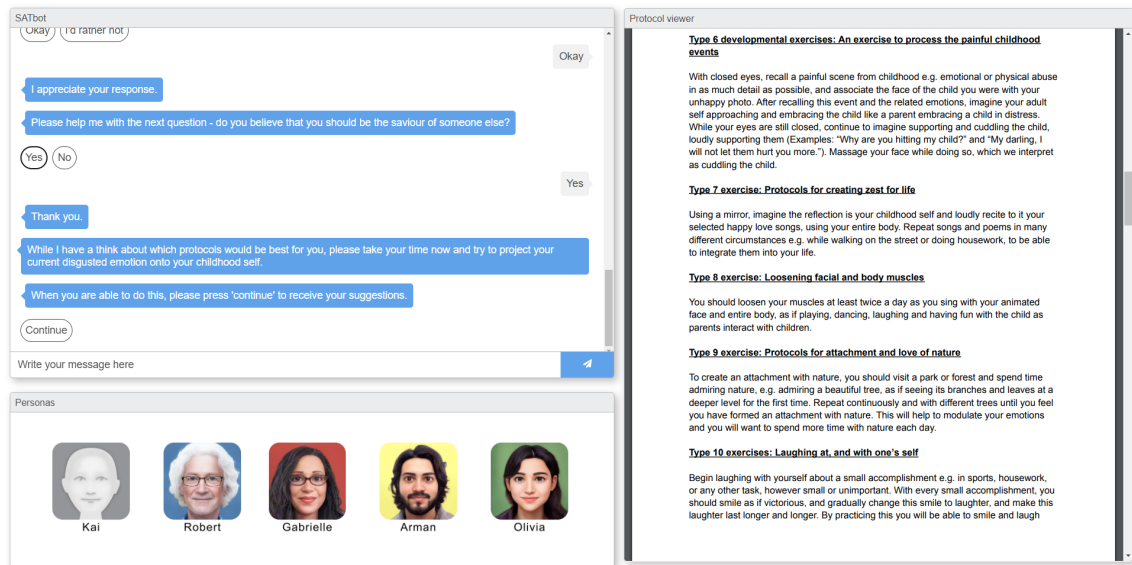


Figure 2.4: User interface of the self-attachment technique chatbot platform [15].

The chatbot initiates a conversation with the user by asking for their name and referring to them personally throughout the dialogue. Before recommending protocols, the chatbot asks the user how they are feeling. This specific prompt and the subsequent answer is classified by the emotion classification model of the chatbot. Emotions are distinguished in two ways for the direction of the discussion - positive or negative. This dictates two different conversation flows and aims to suggest the

²[https:// github.com/LisaAlaz/SATbot](https://github.com/LisaAlaz/SATbot)

most appropriate protocols to the user, according to their mood. During the dialogue the chatbot's utterances are drawn from the augmented EMPATHETICPERSONAS dataset from datapoints that collected for each specific emotion. This has shown to maximise the empathy and engagement of the chatbot [15]. The graphs depicting the two conversation flows can be found in the Appendix.

This existing framework stands as the baseline of this project, which aims to improve the emotion recognition and empathetic response of the chatbot and increase the usefulness of the platform in accordance to self-attachment technique principles. This existing version of the chatbot will be referred to as "Previous Version" in this work.

2.4 Ethical Considerations

As the topic of psychotherapy involves treating mental health problems via different psychological methods, the personal and health information shared by patients to therapists is bound with agreements of non-confidentiality. In the application of virtual psychotherapy, confidential health data is linked to legal and ethical issues.

As described in [1.2 Project Aims](#) human data is collected both at the implementation stage and at evaluation stage of the project. The crowd-sourcing of the dataset and the non-clinical trial are approved by the Research Ethics Committee of Imperial College London. Volunteers providing data are not limited on the nature of information they decide to share, therefore their inputs could contain sensitive data which may fall under the UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018 (DPA) in the UK [55]. Additionally, crowd-sourcing service providers use platform IDs for their workers, which is identifiable information. To comply with the DPA, the collection of the dataset does not include any ID numbers linked to real people and is used entirely for the content in providing human emotional input and response. The handling of any other personal data, such as gender and age is pre-agreed and workers and trial participants are required to consent to how their data is going to be used, stored and processed. Participants are able to request and control the removal of the data they provide.

Furthermore, the virtual chatbot platform does not require identifiable information on the side of participants and it does not store user input, device metadata, network address information or any other data, such as geolocation.

Chapter 3

Data Collection

In the interest of collecting a well-represented and domain-focused dataset for the chatbot, participants were recruited and familiarised with self-attachment therapy. The aim of the crowd-sourcing is to provide sample utterances that would simulate a conversation they would have with the chatbot. The goal is to gather conversations and interactions in different emotional contexts not yet adapted to the chatbot, and increase its emotional corpus. The newly collected data is combined with the available EMPATHETICPERSONAS dataset to create a novel corpus, which is referred to as EMPATHETICPERSONAS12. This section explains the process of how EMPATHETICPERSONAS12 is constructed and analyses the data gathered through crowd-sourcing.

3.1 Process of Data Collection

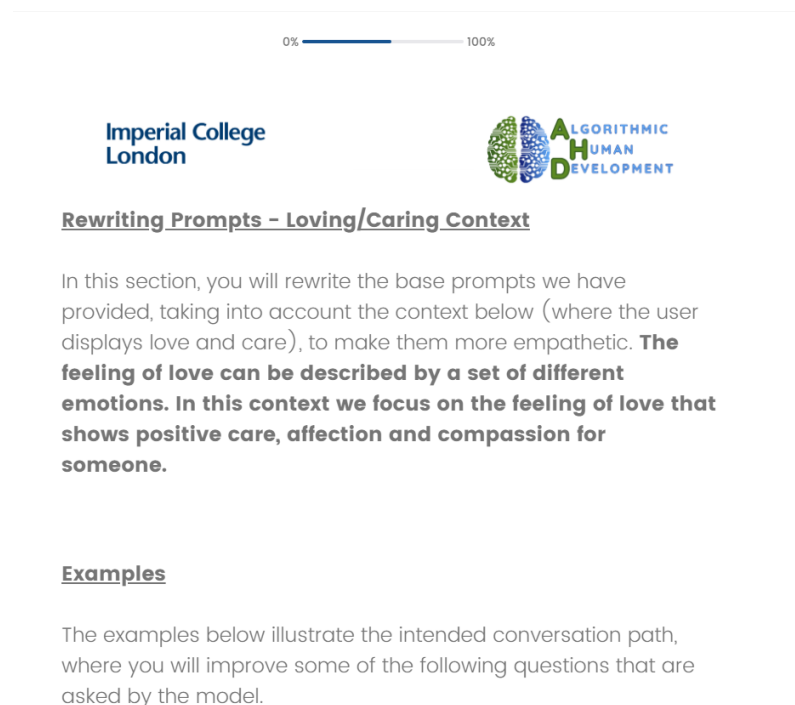
The crowd-sourcing process was conducted in several stages. Firstly, referring to the 32 emotions in the EMPATHETICDIALOGUES dataset [20], eight new emotions were selected in addition to the 4 already present in EMPATHETICPERSONAS [15]. The new emotions were chosen to represent an interest portion of conditions and expressions that are often analysed in attachment theory [56] and subsequently self-attachment theory [12]. The selected new emotions are love, insecurity, disgust, disappointment, shame, guilt, envy and jealousy.

Ambiguous emotions such as anticipation and surprise were deliberately avoided as they can be seen as neutral [19, 57] despite being part of Plutchnik’s 8 proposed basic emotions [16]. As there is no support for neutral emotions in the current version of the virtual chatbot platform these were replaced with love and insecurity. These two emotions capture a wider range of expressions in their definitions, such as care, trust and compassion for love and vulnerability, inferiority and low self-esteem for insecurity. Another design choice is the separate classification of envy and jealousy. As these emotions have been a prevalent part of humans evolutionary psychology their understanding is very important in psychotherapy [58]. Emphasizing the difference between the two emotions, envy is connected to the desire of an outside possession, achievement or quality that one does not have. Jealousy is a feeling expressed towards another another, characterised by fear and lack of trust in the dedication of that person to you. There is currently no known availability of datasets in the field of emotion classification that include both of these emotions, making

EMPATHETICPERSONAS12 a novel corpus that recognises this distinction.

3.2 Survey Creation Process

After selecting the set of the eight new emotions, a total of four surveys were created grouping emotional context in pairs - loving and insecure, disgust and disappointment, shame and guilt, and envy and jealousy. The surveys were designed to capture the conversation flow that a user experiences with the chatbot. The surveys were adapted from previously conducted surveys on the emotions of sadness, anger, anxiety and joy [15]. They were restructured and included with new empathetic prompts and descriptive features of each emotion were added to the context for better understanding.



The screenshot shows a survey interface. At the top, there is a progress bar from 0% to 100%. Below it are the logos for Imperial College London and Algorithmic Human Development. The survey title is 'Rewriting Prompts - Loving/Caring Context'. The main text asks the user to rewrite prompts based on a context of love and care. It includes a section for 'Examples' which explains that the examples illustrate the intended conversation path.

0% 100%

Imperial College London

ALGORITHMIC HUMAN DEVELOPMENT

Rewriting Prompts - Loving/Caring Context

In this section, you will rewrite the base prompts we have provided, taking into account the context below (where the user displays love and care), to make them more empathetic. **The feeling of love can be described by a set of different emotions. In this context we focus on the feeling of love that shows positive care, affection and compassion for someone.**

Examples

The examples below illustrate the intended conversation path, where you will improve some of the following questions that are asked by the model.

Figure 3.1: An example survey focused on the loving emotional context.

3.2.1 Respondent Recruitment

The respondents were recruited on the research platform Prolific¹, which allows for sharing the surveys with a selected group of workers. As per the Prolific Researcher Guidelines², rejection criteria for the survey entries were created. Responses which were less than 50% completed and/or contained low-effort answers, such as copy-pasting the prompts into their answer or included offensive language and profanity were rejected. Based on the target audience of the survey, criteria for the fluency and lexical correctness were carried out at the discretion of the worker's general

¹<https://prolific.co>

²<https://researcher-help.prolific.co/hc/en-gb/articles/360009377834-Approving-Submissions-and-Rejecting-Participants>

ability to respond and understand the survey’s aim. Examples for accepted and rejected answers can be found in the Appendix.

The surveys were released on Prolific in two batches. The surveys for love and insecurity and disgust and disappointment were released first with no target preference for Prolific workers. This allowed for the prompt collection of the results and the time to analyse and process the collected data. However, due to the relative low requirements for the responses, the resulting quality of responses was lower than anticipated. Through feedback received by the workers themselves, the survey was found to be more complicated than anticipated due to its academic jargon. The next batch of surveys on shame and guilt and envy and jealousy were released to an audience that satisfies: native English speakers, living in one of USA, UK, Canada, Republic of Ireland, Australia, New Zealand, South Africa and a completed high-school education as requirement. Resulting responses were improved in quality, however, due to the smaller pool of workers the process took longer.

3.3 Data Analysis

At the end of data collection, the responses included 50 received entries for each survey sent from the first batch and 41 from each survey from the second batch. A total of 993 new emotional and 3419 empathetic utterances were added in EMPATHETICPERSONAS12. Due to the incompleteness of some of the entries, the total collected utterances are broken down in the following way:

	Emotion Samples	Empathy Samples
love	124	47*
insecure	132	510
disgust	122	538
disappointment	126	530
shame	123	440
guilt	121	440
envy	124	457
jealousy	121	457

Table 3.1: Distribution of collected samples for each of the 8 new emotions. *Note that the conversation flow for positive emotions such as "love" is considerably shorter and the number of rewriting required are lower.

3.3.1 Demographic Distribution of the Data

While anonymous, participants consented to provide their age and gender as part of our dataset collection. There were no requirements placed on either study for gender or age. Figures 3.2 and 3.3 present the breakdown in age groups and gender of the given responses. Ages 18-24 completed the highest number of surveys with 91 entries in total, twice as many as the second group of 25-29, which received 46. Following most popular were the ages 30-39 and trailing with just 12 responses in total were 40-49, 60-69, and 50-59, respectively, with 40-49 containing eight entries. This shows that the age demographic was unbalanced with a greater contribution

from the younger popular as 93 % of the responses were from people younger than 40 years old. On the other hand, the gender distribution is comparably more balanced, although female entries tallied at 102, versus 76 for males and five of the responses labelled themselves as non-binary.

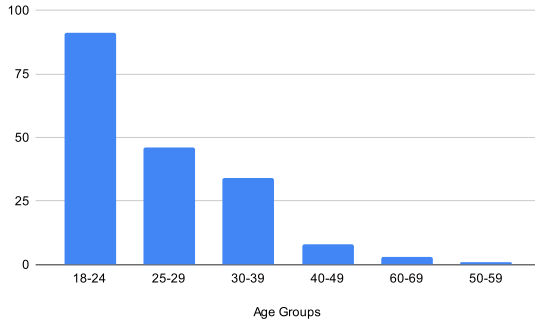


Figure 3.2: Data distribution of age groups

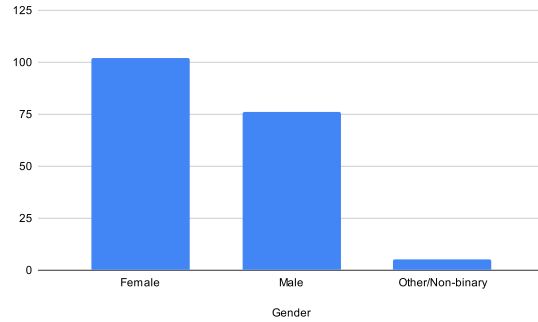


Figure 3.3: Data distribution of gender

3.3.2 Empathetic Utterances

Recruited workers were asked to provide sample rewritings to questions asked by the chatbot in each of the deployed surveys. The utterances were used to create empathetic responses to each emotional context. Following the total collected empathy samples, 3419 new empathetic rewritings were added as datapoints. This is an increase of over 300% compared to the previously sampled rewritings of 1100 in EMPATHETICPERSONAS, which were scored on their empathy discretely from "0" to "2" by annotators [15]. The higher number of samples collected made annotation unachievable for EMPATHETICPERSONAS12. As these newly added samples were coming from the same example questions as the previously annotated utterances, the dataset was labeled in a semi-supervised manner, which is explained in detail in section 4.3.1 Cross-model Semi-supervised Learning Configuration.

As the process of labeling the new empathetic dataset relied on the previously annotated utterances as "ground truth" labels, the new data remains biased towards the initially annotating volunteer's perception of empathy [15]. Thus, it is suggested that in future iterations and usage of EMPATHETICPERSONAS12, empathy data is labeled in a crowd-sourced manner, similar to the collection of EMPATHETICDIALOGUES, where responses were scored on empathy, as well as their relevance and fluency [20]. This would address the need and process of assuring quality of the crowd-sourced utterances by the researcher and remove bias from the labelling, when scoring on aspects such as fluency and novelty as described in 4.3.2 Mutli-objective Optimisation Formula

Chapter 4

Implementation

The main goal of the chatbot is to recommend the most suitable protocol to the patient, based on their emotion. To achieve this the chatbot relies on an internal interpretation of the emotional context of the discussion, which is based on two models trained specifically for the tasks of emotion recognition and empathetic response. This section explores the process of creating these models, including all experiments and their subsequent results, as well as the details about preprocessing and augmenting the EMPATHETICPERSONAS12 corpus.

4.1 Data Augmentation

Due to the differences in crowd-sourcing of EMPATHETICPERSONAS12 the eight new emotions consisted of fewer samples than the four prime ones from EMPATHETICPERSONAS. This created a noticeable class imbalance in the corpus. To avoid model bias and underperformance during training, the dataset was preprocessed and augmented to balance the number of samples for each of the 12 emotions. This process explored several methods for augmenting the data by either downsampling overrepresented emotions or upsampling the underrepresented ones. Upsampling was done using two augmentation techniques - back translation and synonym word replacement.

4.1.1 Downsampling

Downsampling is simply the process of removing datapoints from classes with higher number of datapoints down to the class that has the least datapoints. As shown in Figure 3.1, the minority samples came from "guilt" and "jealousy" at 121. Hence, 121 utterances were sampled randomly from all 12 emotions in EMPATHETICPERSONAS12. This meant cutting quality responses gathered from the collection of the original EMPATHETICPERSONAS corpus.

4.1.2 Back Translation

Back translation is a method that relies on machine translation from a source language to an intermediate language and then back to the source language. In this way, data can be augmented by doubling the sample size while keeping the same sentiment of the utterance but changing the syntax. The technique has been

widely used in research and has proven to be an effective way to upsample text data in the absence of quality labeled data [59]. For performing back translation, the intermediate language of translation was German using a portable version of the Facebook FAIR’s FAIRSEQ model for the WMT19 news translation task submission [60]¹.

4.1.3 Wordnet Synonym Word Replacement

Replacing words with synonyms is another popular way of text data augmentation, which aims to multiply datapoints by choosing random words to be replaced with their dictionary synonyms. For this task, WordNet was chosen as the reference dictionary for the synonyms [61]. WordNet’s lexical database stores words in cognitive synonyms called "synsets" to word’s semantics. However, this happens on a word-level and it does not guarantee the semantical preservation of sentences as is the case of the utterances in EMPATHETICPERSONAS12. During experiments, it was found that replacing more than one word with its WordNet synonym could potentially change the meaning of a sentence. Hence, synonym replacement as a method for upsampling was only done by replacing one word with its WordNet synonym.

4.1.4 Data Augmentation Comparison

Table 4.1 below shows sample runs involving a transformer configuration, fine-tuned on each data augmented corpus of EMPATHETICPERSONAS12 and tested on a held out test set. The choice of the baseline model was ROBERTA_{BASE} in all cases, as its parameter size in comparison to other transformers tested allowed for the efficient training and testing of the data augmentation techniques and choosing the best one.

In addition to upsampling the corpora, the classes were also equalised by sampling the number of datapoints from the minority class. In this case all of the emotions were represented an equal amount of times, and those that were previously over-represented were sampled down.

The best performing configuration was chosen by the highest macro-averaged F1 score on the test dataset. Downsampling the EMPATHETICPERSONAS12 corpus did not perform well, which is expected as it relied on less training data from an already limited corpus. The best configuration during this experiment was the back-translation upsampling technique, which performed better than the equalised version. This dataset was chosen for the emotion classification task and represents all results from models fine-tuned on EMPATHETICPERSONAS12 in the following experiments.

¹<https://huggingface.co/facebook/wmt19-en-de>, <https://huggingface.co/facebook/wmt19-de-en>

Model + Configuration	Accuracy	Macro F1	Weighted F1
ROBERTA _{BASE} + downsampled EP12	74.66	75.50	75.22
RoBERTA_{BASE} + BT upsampled EP12	88.92	89.51	88.79
ROBERTA _{BASE} + WordNet syn. upsampled EP12	88.33	88.67	88.26
ROBERTA _{BASE} + equalised BT EP12	86.60	86.42	86.59
ROBERTA _{BASE} + equalised WordNet syn. EP12	88.32	88.40	88.40

Table 4.1: Comparison of performance between different configurations of ROBERTA_{BASE} fine-tuned on an augmented corpus. All scores are in percentages of a single unit.

Configuration	Accuracy	Macro F1	Weighted F1
Downsampled EP12	74.66	75.50	75.22
BT upsampled EP12	88.92	89.51	88.79
WordNet syn. upsampled EP12	88.33	88.67	88.26
Equalised BT EP12	86.60	86.42	86.59
Equalised WordNet syn. EP12	88.32	88.40	88.40

Table 4.2: Comparison of performance between different configurations of ROBERTA_{BASE} fine-tuned on an augmented corpus. All scores are in percentages of a single unit.

4.2 Emotion Classification

As established by the previous version of the chatbot, T5 and ROBERTA transformers performed significantly well on the emotion classification task, fine-tuned on EMPATHETICPERSONAS [15]. In addition to implementing ROBERTA_{BASE} and T5_{BASE} for emotion classification, ROBERTA_{LARGE} and MEGATRON-BERT were also implemented and trained on EMPATHETICPERSONAS12. The configuration with the highest macro-f1 score was chosen to be implemented as the emotion classification model in the chatbot application.

4.2.1 Fine-Tuning

Fine-tuning a transformer model is a technique that involves using a large language model such as the ones mentioned in [Transformers](#) by loading a pre-trained checkpoint. The training of these models takes a lot of computation power and a large amount of data, which needs to be processed and implemented into the transformer’s architecture. In order to use transformers for a custom downstream task such as emotion multi-class classification, configuring a custom training point and then using a specific dataset allows transformers to solve any custom downstream task.

Fine-Tuning on EMPATHETICPERSONAS12

During experiments, all of the transformers were trained on the EMPATHETICPERSONAS12 corpus. The utterances from the dataset are tokenised before being sent as an input to the models. BERT-based models utilise a byte-level byte-pair encoding [\[62\]](#) tokeniser similarly to GPT-2 [\[39\]](#) while T5 uses SentencePiece [\[63\]](#) as its tokenisation algorithm [\[48\]](#). Both of these are subword-based, meaning they tokenise both at the word-level and at the character-level and keep a dictionary of known and encoded words [\[64\]](#) ².

Double fine-tuning on EMPATHETICDIALOGUES and EMPATHETICPERSONAS12

EMPATHETICDIALOGUES contains a large volume of utterances and compared to other datasets such as GOEMOTIONS [\[19\]](#), it has the highest amount of labelled emotions that are relevant to the targeted ones in the EMPATHETICPERSONAS12. The dialogue-based datapoints also hold similar sentiment to the utterances expected by the chatbot. Unfortunately, EMPATHETICDIALOGUES does not have distinctive labels for envy and jealousy, but the "jealous" datapoints given in EMPATHETICDIALOGUES were found to be conforming to the definition of "envy" more so than their recognised label. Hence, when double fine-tuning on both EMPATHETICDIALOGUES and EMPATHETICPERSONAS12 jealousy-labeled utterances were ignored. As a result, 11 emotions were matched from the EMPATHETICDIALOGUES dataset to the crowd-sourced corpus. ³

²In order to make the labels of the data inputs from EMPATHETICPERSONAS12 encoded as single token ids, some emotions were changed, as the words were not in T5’s dictionary. "Insecurity" was changed to "Instability" and "Jealousy" changed to "jealous". This modification is prevalent in every experiment for emotion classification.

³The 11 labels between the two datasets were matched in the following manner (ED to EP12): sad to sadness; angry to anger; joyful to joy; afraid to fear; caring to love; apprehensive to insecurity; disgusted to disgust; disappointed to disappointment; ashamed to shame; guilty to guilt; jealous to envy.

4.2.2 Model Evaluation

Model + Configuration	Accuracy	Macro F1	Weighted F1
ROBERTA _{BASE} + EP12	88.92	89.51	88.79
T5 _{BASE} + EP12	80.06	80.13	80.07
ROBERTA _{LARGE} + EP12	89.24	89.40	89.16
MEGATRON-BERT + EP12	89.56	90.06	89.55
ROBERTA _{BASE} + ED + EP11*	91.25	91.20	91.27
T5 _{BASE} + ED + EP11*	86.31	85.96	86.25
ROBERTA _{LARGE} + ED + EP11*	91.25	91.28	91.16
MEGATRON-BERT + ED + EP11*	92.02	91.79	92.03

Table 4.3: Comparison of performance between different configurations on the emotion classification task. All scores are in percentages of a single unit. *Note that given the unavailability of different jealousy and envy labels in EMPATHETICDIALOGUES, the used emotions from the EMPATHETICPERSONAS12 corpus are now 11.

A total of eight different configurations were tested for their performance and evaluated on a held out test set. In each setting, the corpora were split in a ratio of 80:10:10 for train, validation and test datasets, respectively.

The different models were compared based on their macro-averaged F1 scores and the best model was chosen to be one scoring highest in that component. Out of the experiments run MEGATRON-BERT + ED + EP12 proved to be highest scoring configuration on the task at 91.79 % (Table 4.3).

4.3 Empathy Classification

As EMPATHETICPERSONAS12 adds 3419 new unlabeled empathetic datapoints, a 300% increase from the original 1100 utterances, the labelling of this data is specifically complex due to the resources needed to annotate all datapoints. Instead, EMPATHETICPERSONAS12’s empathy is labelled in a semi-supervised manner by combining the 1100 already labelled datapoints and the newly collected ones to create a weakly-labelled corpus of 4519 empathetic utterances. Annotator-provided labels from EMPATHETICPERSONAS will be referred to as "true" labels during this process.

4.3.1 Cross-model Semi-supervised Learning Configuration

Semi-supervised learning is a common approach of dealing with a corpus that contains few true labels in problems where high-quality labelled data is expensive, such as medicine or law [65]. As the rewritings in the weakly-labelled dataset are based on the same questions, this classifies the corpus as one-domain specific

dataset. Self-training is a semi-supervised learning approach that leverages such a weakly-labelled dataset and uses a teacher model, which creates synthetic labels by fine-tuning on the ground-truth ones [66]. A student model is fine-tuned on the datapoints of true and synthetic labels and evaluated on a held out test data. In the case of self-training, the teacher model is a smaller model in terms of parameters than the student one, whereas a larger teacher model teaching a smaller student one is classified as knowledge distillation [66].

Training on the weakly-labeled empathetic corpus is done in a two-way experiment with self-training and knowledge distillation through a cross-model network. The cross-model T5 and ROBERTA configuration is used in order to perform an unbiased evaluation of the retrieval of each others' empathy scoring. Figure 4.1 shows the self-training and knowledge distillation setups used for the empathetic scoring. The labels were consistent with the ones annotators used in EMPATHETICPERSONAS and all metrics were evaluated based on a held out true-labeled test dataset that has not been seen by either model.

When evaluating the self-training and knowledge distillation tasks, the final metrics are compared against the achieved results in empathy classification only on the annotated dataset. Self-training shows an increase in the accuracy and macro-averaged F1 score compared to second best model. This is an improvement over the empathy classification model used in the previous version of the chatbot, shown by T5 trained on the true-labelled dataset in Figure 4.4.

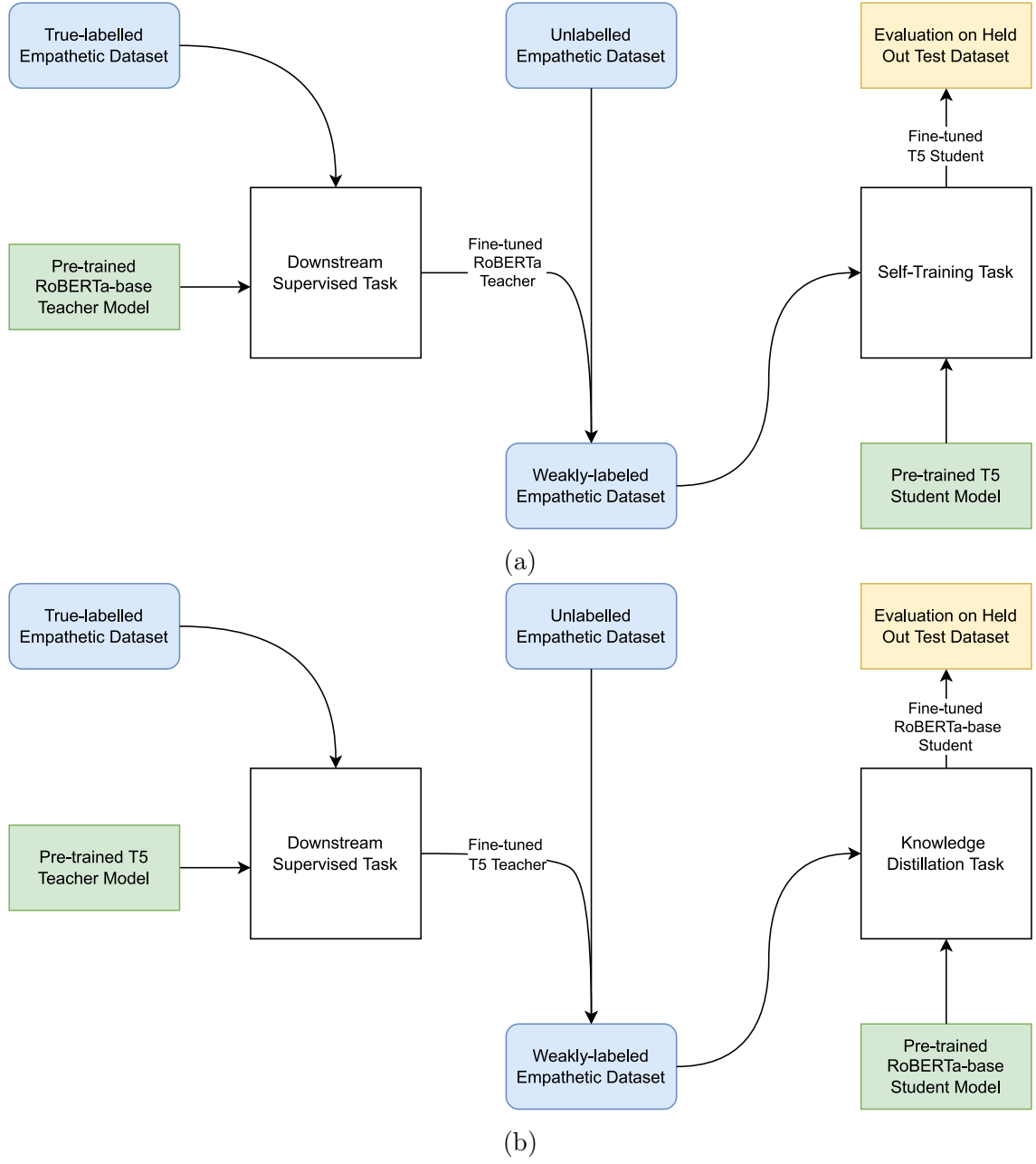


Figure 4.1: Diagram of cross-model pipeline for (a) self-training between a RoBERTa-base teacher and a T5 student model (b) knowledge distillation between a T5 teacher model and a RoBERTa-base student model on the weakly-labelled empathetic dataset.

Training Process	Accuracy	Macro F1
RoBERTA _{BASE} on true-labelled data	73.87	74.72
T5 on true-labelled data	80.18	80.66
Self-Training (RoBERTA_{BASE} teacher and T5 student)	81.08	81.36
Knowledge Distillation (T5 teacher and RoBERTA _{BASE} student)	77.48	77.99

Table 4.4: Comparison of performance between models trained on true-labelled data and semi-supervised trained ones. All scores are in percentages of a single unit.

4.3.2 Mutli-objective Optimisation Formula

Outside of the empathy scoring achieved by self-training, fluency and novelty of the response are also considered in a dialogue with the virtual chatbot. The evaluation and selection of a response, based on the user’s expression are computed during the runtime of the platform and leverage a multi-objective function, as proposed in the previous version of the chatbot [15].

$$R(u) = w_e E_{norm}(u) + w_f F_{norm}(u) + w_d D_{norm}(u) \quad (1)$$

In (1), $w_e E_{norm}(u)$, $w_f F_{norm}(u)$, $w_d D_{norm}(u)$ are denoted to be the weighted normalised functions measuring the empathy score, fluency and novelty, respectively, of the utterance u [15]. Optimising $R(u)$ was tested comprehensively, by tuning the weights of each function respectively. Notably, fluency score is calculated by $F(u) = \frac{1}{\text{PPL}(u)} - \text{RP}(u)$, where the perplexity $\text{PPL}(u)$ is computed by GPT-2 and $\text{RP}(u)$ is a penalty score that is subtracted from the inverse perplexity in the calculation.

The penalty score is dependent on the repetition of each token in an utterance with the exception of stop words. During the collection of EMPATHETICPERSONAS12, many acceptable entries contained repetitions of single words. This could be explained by the difference in recruiting the workers in the collection of the corpus, compared to EMPATHETICPERSONAS. This is not seen as an essential disadvantage in the responses of the chatbot. Hence, the previously tuned penalty of 10^{-2} was adjusted to be 0.5×10^{-2} in order to prevent the chatbot from discriminating against responses collected and augmented during the course of crowd-sourcing EMPATHETICPERSONAS12. To address the increase in normalised fluency, the weights in (1) have been adjusted to $w_e = 1$, $w_f = 0.675$, $w_d = 1$, from previously set $w_e = 1$, $w_f = 0.75$, $w_d = 1$ [15], as this was found to have the best performance in retrieving a variety of responses, based on the quality of data crowd-sourced for the eight new emotions.

4.4 Hyperparameter Tuning

Choosing the hyperparameters for fine-tuning each model to the dataset was done by tuning on four parameters: batch size, learning rate, accumulating gradient batches and training epochs. All parameters were evaluated using a grid search algorithm

across several possible values. The metric for this search was to minimise the cross-entropy loss ⁴ during the learning process. The optimiser used is AdamW ⁵ with no weight decay and an epsilon of $\epsilon = 1^{-6}$. The table below shows all parameters searched for the hyperparameter process and the ones that were chosen for best performance and used for the all models outlined in this section.

Parameter	Range	Best
Learning Rate	$[1.35 \times 10^{-4}, 5 \times 10^{-4}, 1.35 \times 10^{-5}, 5 \times 10^{-5}, 1.35 \times 10^{-6}]$	5×10^{-5}
Batch Size	[4,8,16,32]	16
Accumulated Gradient Batches	[1,2,4,8]	2
Epochs	[6,10,16,20]	10

Table 4.5: Parameters selected for tuning and the best performing combination for the classification tasks.

⁴<https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html>

⁵https://huggingface.co/docs/transformers/main_classes/optimizer_scheduletransformers.AdamW

Chapter 5

Non-clinical trial

In order to evaluate the performance of the newly created virtual agent and the extent of its functionality, a non-clinical trial was conducted. Volunteers were recruited to interact with the platform and were familiarised with the self-attachment therapy protocols and the scope of the study. A total of 16 users took part in the study, completing a post-trial survey, the data of which we analyse in this section.

5.1 Trial Setup and Limitations

The trial’s aim was to measure the qualitative performance of the platform and the overall engagement with the virtual agent, as well its performance on emotion recognition and empathetic response. While constructing the trial and the post-trial survey, previously measured evaluations were consulted to create a base for comparison [15]. In order to keep an unbiased experience, the volunteers were instructed on which emotions the chatbot is able to recognise (12 in total) and were asked to interact with the agent twice, by choosing two different emotional contexts. They were instructed on how to use the platform before the trial and were not invigilated for the purpose of not influencing the efficacy of the results.

In this trial, users evaluated the fluency, engagement and usefulness of the platform, as well as the correctness of the emotion recognition and the empathy of the model’s response by completing a Likert Scale questionnaire. They were also asked to identify which emotions the chatbot recognised, and what it misinterpreted them as, if they were not recognised. These questions were used in estimating the model’s performance based on user experience as well as record what emotional contexts were most prevalent overall in the engagement with the chatbot.

In the previously run trial the platform offered the choice of interacting with 5 personas, namely Kai, Robert, Gabrielle, Arman and Olivia. Kai combines all responses from the EMPATHETICPERSONAS corpus, while the other 4 personas were modeled based on gender and age of the collected responses [15]. For this trial, however, limitations in collecting a well-represented empathetic prevented the modelling of the 4 diversified personas to the new version of the chatbot. Participants were instructed to interact only with Kai during the trial.

5.2 Trial Results

When it came to scoring the platform on engagement, fluency and usefulness the results were overall favourable. Out of the 16 participants in the trial, all agreed that the chatbot displayed proper lexical and grammatical language and 75% of the volunteers found the platform at least somewhat useful, while 25% were neutral about its usefulness. The engagement with the chatbot, in this case Kai, also saw commendatory results, however, three of the trial participants disagreed that the conversation was engaging for them with one result showing a strong disagreement. Nevertheless, 10 participants, more than half of all taking part, found the conversation with Kai somewhat or strongly engaging. The distribution of these statistics is shown in Figure 5.1.

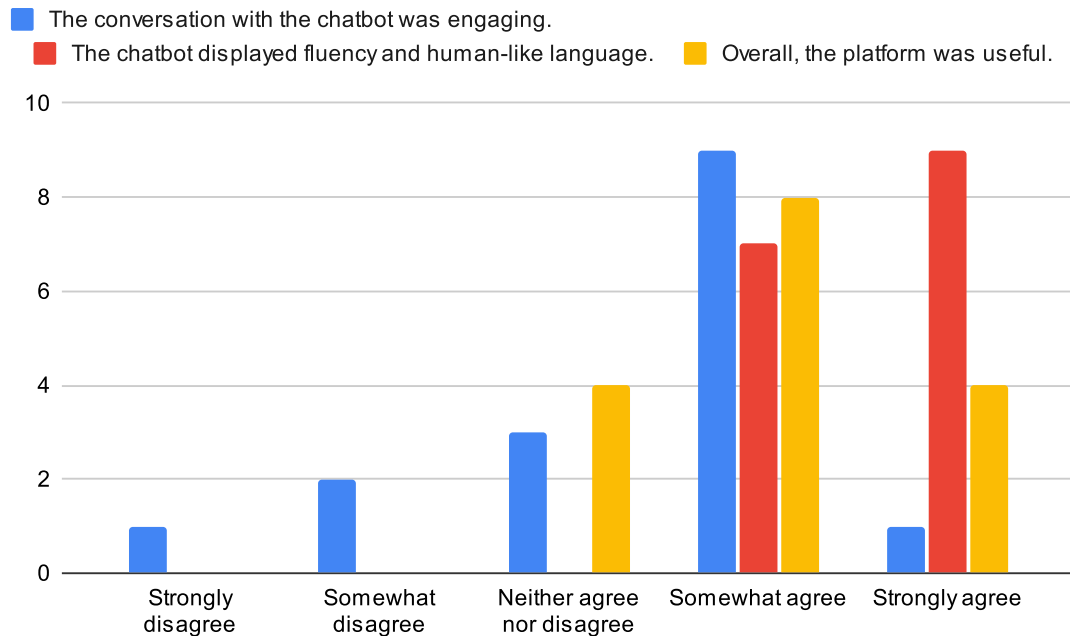


Figure 5.1: Distribution of trial responses across engagement, fluency and overall usefulness.

The model's emotion and empathy classification was evaluated by whether the chatbot guessed the user's emotion and whether it conversed with empathy to the conversation. The findings in Figure 5.2 show that 75% of the participants strongly or somewhat agree that the chatbot guessed their emotions. 12.5% of the users were neutral about the chatbot recognising their emotion and another 12.5% said that the chatbot failed at recognising their emotions. The empathy results in Figure 5.3 show that over 68% of the responses agree that the chatbot reacted with empathy to their emotions, while a quarter replied with neutrality about the chatbot's empathy and one participant disagreed with the statement. This shows that, while there were mislabelling errors on the side of the emotion classification model, the chatbot's emotion recognition was received more favourably, compared to the empathetic response.

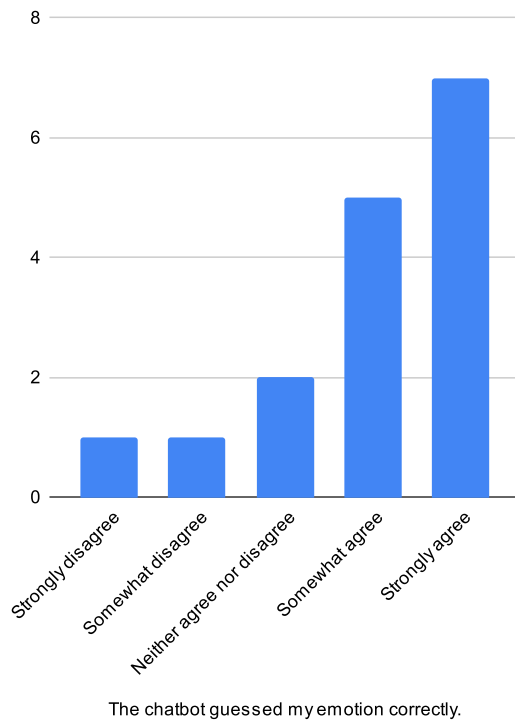


Figure 5.2: Distribution of responses on the chatbot's emotion recognition.

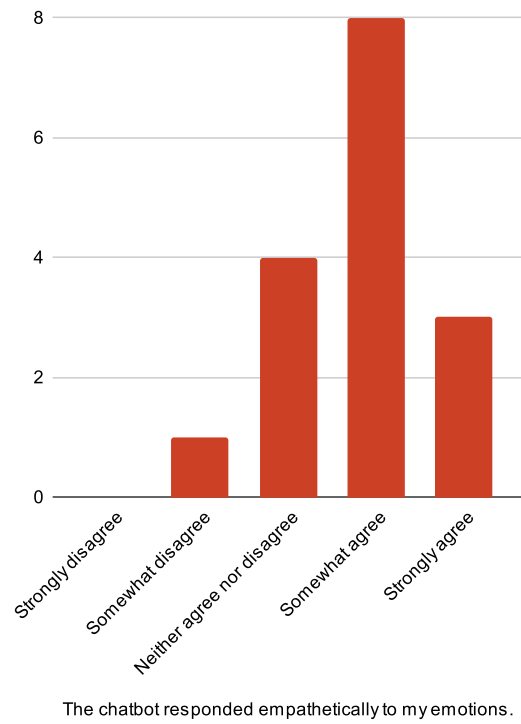


Figure 5.3: Distribution of responses on the chatbot's empathy.

The most popular emotion expressed to the chatbot was anxious/fearful at 11 interactions, followed by happy/content at 5, guilt at 4, and love at 2. Disappointment, shame, and envy were recorded to be expressed 1 time each (Figure 5.4).

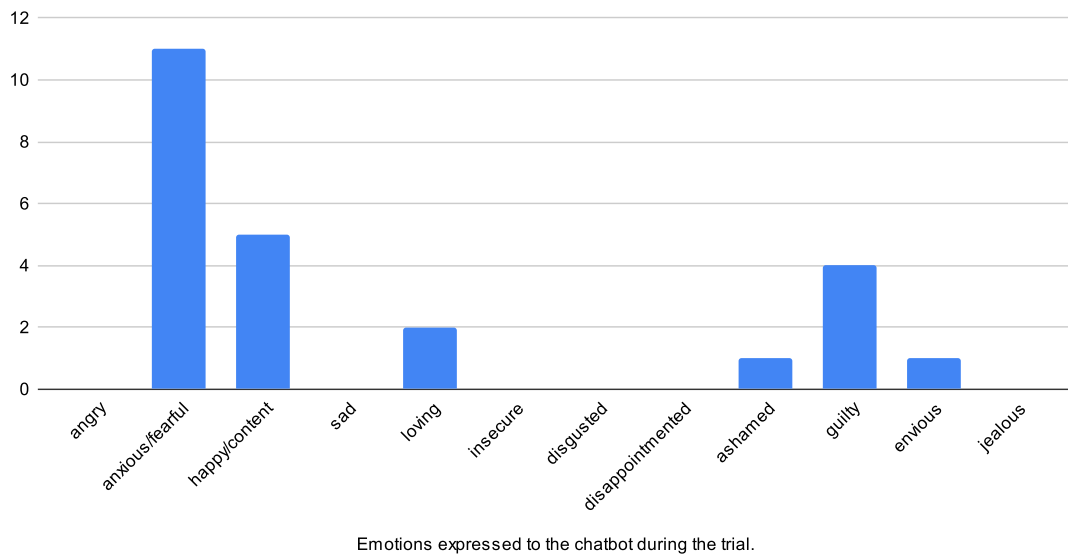


Figure 5.4: Distribution of emotions expressed by participants in the trial.

Chapter 6

Project Evaluation

Following the completion of the implementation and the subsequent non-clinical trial, the project has achieved the aims of increasing the emotional recognition and empathy capacity of the virtual self-attachment therapy chatbot. The outcomes of this project are evaluated by comparing with the results of the previous version of the chatbot and its achievements.

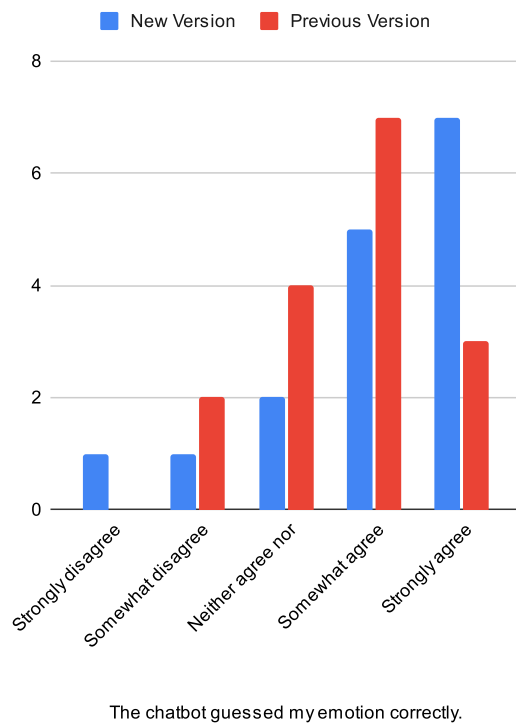


Figure 6.1: Comparison of trial responses rating the chatbot's emotion recognition.

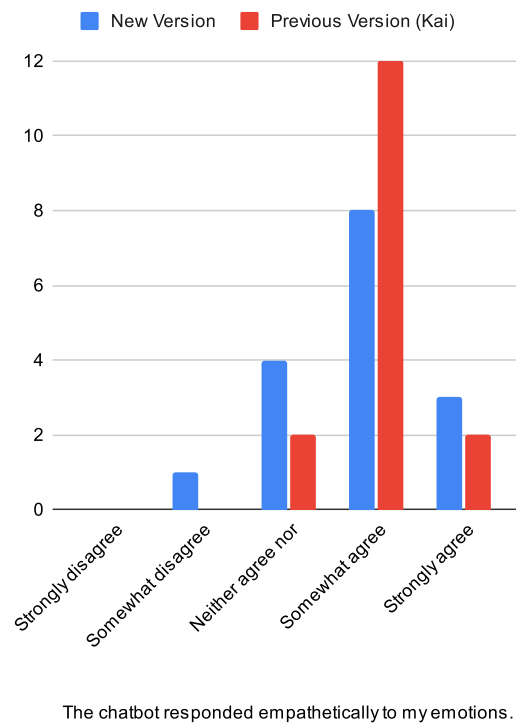


Figure 6.2: Comparison of trial responses rating the chatbot's empathy.

Both the previous and the new version of the chatbot was assessed in a non-clinical trial of 16 people. When comparing user satisfaction with the emotional and empathetic abilities of the chatbot, the new version sees a slight improvement in how the chatbot guessed their emotion. The empathy of the chatbot did not see any improvements over the previously implemented version, with considerably

fewer participants favouring the empathy classification of the new chatbot’s model, compared to the old one. While recognising empathy is subjective, these results could be explained from the relatively little to no customisation of the responses of the chatbot to specific emotions, considering there are 3 times as many in the new dataset. The discussion is swayed towards negative emotions as a whole and the feelings of some of the new more complex emotions are not targeted in a special empathetic manner. Figures 6.1 and 6.2 show the trial results for assessing the chatbot’s ability emotional and empathetic abilities.

Trials on the engagement of the chatbot show a slight decrease in engagement from previous versions. Engagement was positively accepted by 62.5% of the participants, slightly lower than the previous trial, but a larger proportion of people found the chatbot to be non-engaging. Despite comparing both results from discussion with the persona Kai, a potential improvement of engagement can come from tailoring the new emotions and empathetic responses to the previously implemented personas. Targeted data collection in the direction of creating specific range of personas could be one way to increase the datapoint necessary for more diverse personas or implementing data augmentation to the current EMPATHETICPERSONAS12 dataset in order to upsample the undersampled datapoints.

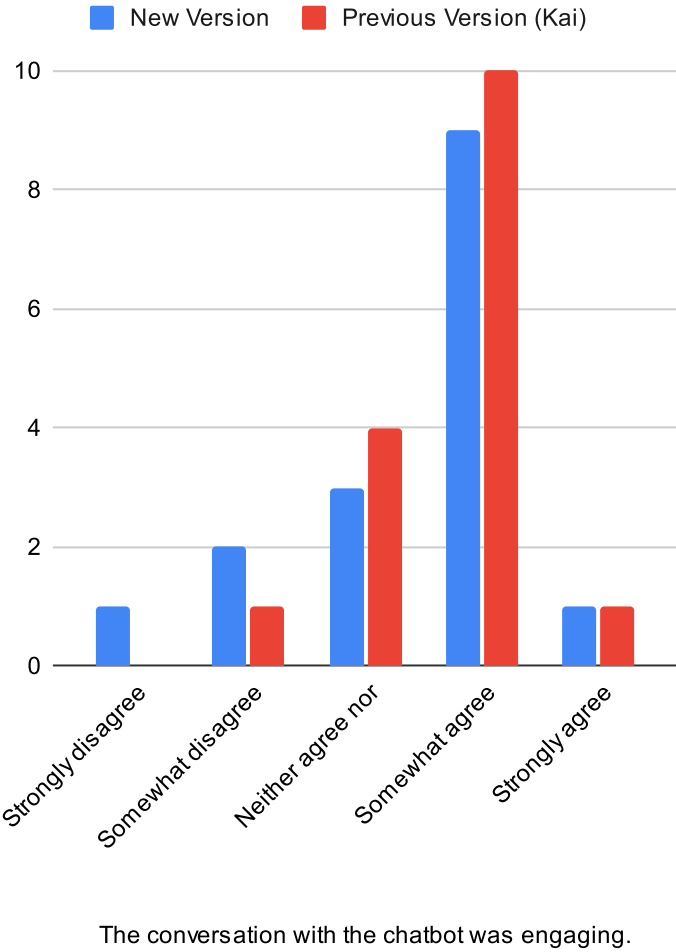


Figure 6.3: Comparison of trial responses assessing the platform’s engagement.

Usefulness of the platform was received neutrally as well. While none of the participants showed disagreement with the statement, fewer people, at 75% of respondents, found the platform useful, compared to 93.75% from the previous trial. This metric could be influenced by the environment in which the study was done. While the previous trial was conducted remotely, the current one was done in person. Participants did not have access to the platform at their own convenience, but rather a timetabled time of the day. This could prevent users from the conditions in which they can confidently and calmly practice self-attachment technique protocols. A suggested improvement in this category is to have the chatbot introduce the concepts and aims of the protocols in order to facilitate the user’s expectations and knowledge of the exercise.

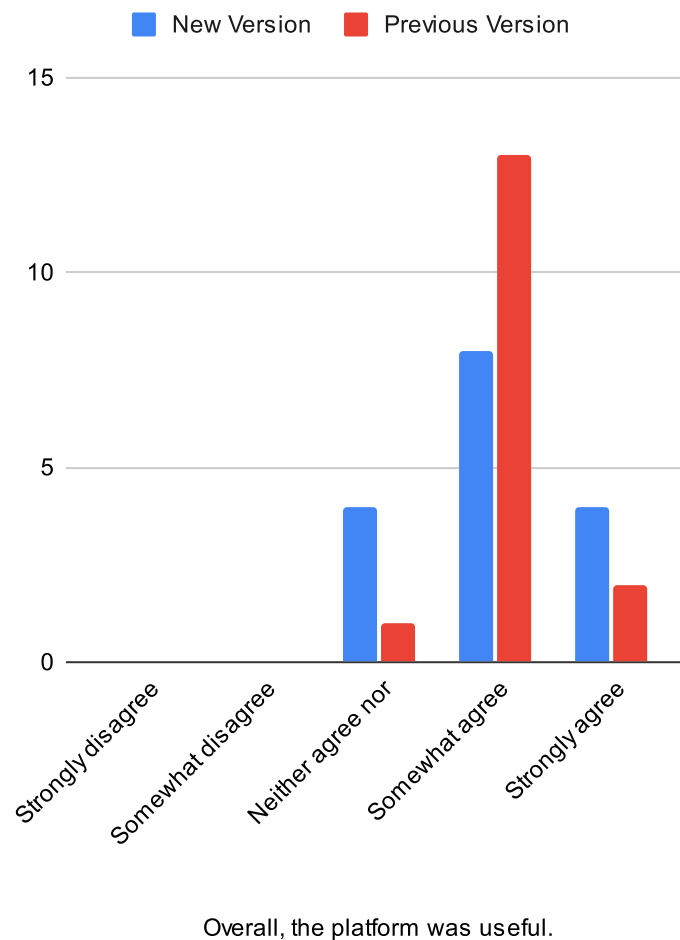


Figure 6.4: Comparison of trial responses assessing the platform’s usefulness.

The performance comparison between the two implemented models for the chatbot’s classification tasks could not be objectively assessed, due to different factors impacting the performance. Fundamentally, the EMPATHETICPERSONAS12 corpus is now a larger dataset than the previous version of the chatbot trained on EMPATHETICPERSONAS and contains 3 times as many emotional contexts. The datapoints cannot be claimed to be specifically fine-grained to the emotions described - they were provided by recruited users, who had their own subjective interpretation of the emotions. As such, emotional contexts in the corpus were

prone to human misinterpretation and subsequently model interpretation as well. This is especially the case for emotions who are similar in contexts, such as sadness and insecurity, joy and love, shame and guilt, and jealousy and envy.

Despite that, this work claims that the model performance is of satisfying quality, when compared to other popular work in the field of emotion classification and this is supported by the outcome of the non-clinical trials, which showed a confidently favourable receipt of the chatbot’s emotion recognition abilities.

There is also an observed improvement in the empathy classification scores owing to a successful self-supervised training of an initially weakly-labelled dataset. The larger corpus of the same domain and class system, contributes to a larger amount of datapoints for the model to improve on. Despite this, a downside in the empathy classification training task is the potential lack of objectivity in recognising and labelling the empathetic responses. As with the previously implemented version of the chatbot trained on EMPATHETICPERSONAS, this iteration did not crowd-source the empathy scoring of the used corpus and relied on the labelling of the initially selected annotators.

	Emotion Classification	Empathy Classification
Previous Version	95.10	80.66
New Version	91.79	81.36

Table 6.1: Comparison of the macro-averaged F1 scores for the best models in emotion and empathy classification between the previous and the new version of the chatbot. All scores are in percentages of a single unit.

Chapter 7

Conclusion and Future Work

This work shows the benefits of developments in natural language processing when combined with self-attachment theory as a form of virtual therapy. While the evaluation of the project shows positive achievements it also outlines several factors to focus on when improving the virtual self-attachment therapy chatbot.

The non-clinical trial was a beneficial stage for the evaluation of the project. For more objective results, the trial should be carefully moderated and volunteers should be familiarised with the concept of self-attachment therapy and its protocols. It was conducted in person without asking participants about their preference. Whether the trial is done in person at specific times during the period or at the discretion of their own time and need is highly influential in how they would evaluate the platform and the experience. Unlike the previous version of the chatbot, the new one did not receive feedback from trained professionals during the non-clinical trial. Specialised areas such as psychotherapy and self-attachment therapy in particular would benefit from the assessment of experienced researchers. At its current version the chatbot is not tested by specialists and is not known how it would perform in a clinical therapy session.

Despite successfully collecting a well-represented emotion corpus, EMPATHETICPERSONAS12 lacks demographically diverse data. This led to the unavailability of other personas in the new version of the chatbot, which has impacts on the engagement and usefulness of the platform. Obstacles in crowd-sourcing a quality emotion dataset also include the subjectivity of how one defines a certain emotion. This problem is only amplified when an increase in the number of emotions in the corpus is observed. Future work on the self-attachment chatbot should focus on improving the EMPATHETICPERSONAS12 corpus by forming multi-labels in utterances that are of similar sentiment. This process can also be crowd-sourced in order to avoid bias coming from a small number of annotators. Examples of multi-label emotion datasets already exist, such as GOEMOTIONS[19, 57] and this dataset can also be implemented in fine-tuning the models for solving a multi-label classification problem. Additionally, scoring empathetic data currently relies on manually labelled samples from a limited number of annotators. Crowd-sourcing the labelling of empathetic utterances should be considered in future iterations in order to improve the objectivity of results in empathy classification.

The training process for the transformers can also be improved when working towards an optimal model for emotion recognition. More sophisticated methods

include training a custom network that prepends the top predicted label of an utterance to the beginning and using it as an encoder input. This is proposed by Rashkin et al. during experiments on the EMPATHETICDIALOGUES dataset [20].

When it comes to the existing chatbot, the flow of conversation could be improved to suggest specific protocols that target a group of emotions. There are currently 10 negative emotions recognised by the chatbot that follow the same dialogue paths. The framework could be also adapted offer descriptions of how the suggested protocol would help the specific emotion felt by the patient. This could increase the interest and motivation for the user to apply the protocols and increase their mindfulness during the exercise.

Bibliography

- [1] WHO. Mental health;. https://www.who.int/health-topics/mental-health#tab=tab_2.
- [2] Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*. 2020;395(10227):912-20. Available from: <https://www.sciencedirect.com/science/article/pii/S0140673620304608>.
- [3] Sprang G, Silman M. Posttraumatic Stress Disorder in Parents and Youth After Health-Related Disasters. *Disaster Medicine and Public Health Preparedness*. 2013;7(1):105–110.
- [4] Minde K. Affect Dysregulation and Disorders of the Self. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*. 2006 May;15(2):100-1. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2277284/>.
- [5] MIKULINCER M, SHAVER PR. An attachment perspective on psychopathology. *World Psychiatry*. 2012 Feb;11(1):11-5. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3266769/>.
- [6] Rimmer A. Mental health: Staff shortages are causing distressingly long waits for treatment, college warns. *BMJ*. 2021;375. Available from: <https://www.bmj.com/content/375/bmj.n2439>.
- [7] Wainberg ML, Scorza P, Shultz JM, Helpman L, Mootz JJ, Johnson KA, et al. Challenges and Opportunities in Global Mental Health: a Research-to-Practice Perspective. *Current Psychiatry Reports*. 2017 May;19(5):28. Available from: <http://link.springer.com/10.1007/s11920-017-0780-z>.
- [8] Weightman M. Digital psychotherapy as an effective and timely treatment option for depression and anxiety disorders: Implications for rural and remote practice. *Journal of International Medical Research*. 2020;48(6):0300060520928686. PMID: 32527170. Available from: <https://doi.org/10.1177/0300060520928686>.
- [9] Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Mental Health*. 2019 Oct;6(10):e14166. Available from: <https://mental.jmir.org/2019/10/e14166>.

- [10] Hofmann SG, Asnaani A, Vonk IJJ, Sawyer AT, Fang A. The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive Therapy and Research*. 2012 Oct;36(5):427-40. Available from: <http://link.springer.com/10.1007/s10608-012-9476-1>.
- [11] Edalat A. Self-attachment: A self-administrable intervention for chronic anxiety and depression; 2016. <https://www.doc.ic.ac.uk/research/technicalreports/2017/DTRS17-3.pdf>.
- [12] Edalat A. Self-attachment: A holistic approach to Computational Psychiatry; 2017. <https://www.doc.ic.ac.uk/~ae/papers/self-attachment-f.pdf>.
- [13] Edalat A, Farsinezhad M, Bokharaei M, Judy F. A Pilot Study to Evaluate the Efficacy of Self-Attachment to Treat Chronic Anxiety and/or Depression in Iranian Women. *International Journal of Environmental Research and Public Health*. 2022;19(11). Available from: <https://www.mdpi.com/1660-4601/19/11/6376>.
- [14] Ghachem A. Evaluation of a Virtual Agent in Guiding Users from the Non-Clinical Population in Self-Attachment Intervention; 2021. https://www.doc.ic.ac.uk/~ae/papers/Ali_Ghachem_report.pdf.
- [15] Alazraki L. A deep-learning assisted empathetic guide for self-attachment therapy; 2020. https://www.doc.ic.ac.uk/~ae/papers/Lisa_Alazraki_report.pdf.
- [16] Ekman P. An argument for basic emotions. *Cognition and Emotion*. 1992;6(3-4):169-200. Available from: <https://doi.org/10.1080/02699939208411068>.
- [17] Plutchik R. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*. 2001;89(4):344-50. Available from: <http://www.jstor.org/stable/27857503>.
- [18] Saravia E, Liu HCT, Huang YH, Wu J, Chen YS. CARER: Contextualized Affect Representations for Emotion Recognition. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3687-97. Available from: <https://aclanthology.org/D18-1404>.
- [19] Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: A Dataset of Fine-Grained Emotions. In: *58th Annual Meeting of the Association for Computational Linguistics (ACL)*; 2020. .
- [20] Rashkin H, Smith EM, Li M, Boureau Y. I Know the Feeling: Learning to Converse with Empathy. *CoRR*. 2018;abs/1811.00207. Available from: <http://arxiv.org/abs/1811.00207>.
- [21] Rogers CR, Koch S. *A Theory of Therapy, Personality, and Interpersonal Relationships: As Developed in the Client-centered Framework*. McGraw-Hill; 1959. Available from: <https://books.google.co.uk/books?id=zsIBtwAACAAJ>.

- [22] Rogers CR, Stevens B, Gendlin ET, Shlien JM, Van Dusen W. Person to Person: the Problem of Being Human: A New Trend in Psychology. Condor books. Real People Press; 1967. Available from: <https://books.google.co.uk/books?id=-UweAQAAIAAJ>.
- [23] Sharma A, Miner A, Atkins D, Althoff T. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 5263-76. Available from: <https://aclanthology.org/2020.emnlp-main.425>.
- [24] Rashkin H, Smith EM, Li M, Boureau YL. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset; 2019.
- [25] Levinson W, Gorawara-Bhat R, Lamb J. A Study of Patient Clues and Physician Responses in Primary Care and Surgical Settings. JAMA. 2000 08;284(8):1021-7. Available from: <https://doi.org/10.1001/jama.284.8.1021>.
- [26] Bickmore T, Cassell J. Relational Agents: A Model and Implementation of Building User Trust. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '01. New York, NY, USA: Association for Computing Machinery; 2001. p. 396–403. Available from: <https://doi.org/10.1145/365024.365304>.
- [27] Zhong P, Zhang C, Wang H, Liu Y, Miao C. Towards Persona-Based Empathetic Conversational Models; 2020. p. 6556-66.
- [28] Brave S, Nass C, Hutchinson K. Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. International Journal of Human-Computer Studies. 2005 02;62:161-78.
- [29] Wright P, McCarthy J. Empathy and Experience in HCI. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '08. New York, NY, USA: Association for Computing Machinery; 2008. p. 637–646. Available from: <https://doi.org/10.1145/1357054.1357156>.
- [30] Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health. 2017 Jun;4(2):e19. Available from: <http://mental.jmir.org/2017/2/e19/>.
- [31] Scott D. Some definitional suggestions for automata theory. Journal of Computer and System Sciences. 1967;1(2):187-212.
- [32] Chaitin GJ. Algorithmic information theory. IBM journal of research and development. 1977;21(4):350-9.

- [33] Su KY, Chiang TH, Chang JS. An overview of corpus-based statistics-oriented (CBSO) techniques for natural language processing. *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 1, Number 1, August 1996. 1996:101-58.
- [34] Jurafsky D, Martin JH. *Speech and Language Processing*; 2021. <https://web.stanford.edu/~jurafsky/slp3/>.
- [35] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. *CoRR*. 2017;abs/1706.03762. Available from: <http://arxiv.org/abs/1706.03762>.
- [36] Kombrink S, Mikolov T, Karafiát M, Burget L. Recurrent Neural Network Based Language Modeling in Meeting Recognition.; 2011. p. 2877-80.
- [37] Wang C, Li M, Smola AJ. Language Models with Transformers. *CoRR*. 2019;abs/1904.09408. Available from: <http://arxiv.org/abs/1904.09408>.
- [38] Mazaré PE, Humeau S, Raison M, Bordes A. Training Millions of Personalized Dialogue Agents. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 2775-9. Available from: <https://aclanthology.org/D18-1298>.
- [39] Alec Radford RCDLDAIS Jeffrey Wu. Language Models are Unsupervised Multitask Learners; 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- [40] Better Language Models and Their Implications; 2019. <https://openai.com/blog/better-language-models/>.
- [41] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. *CoRR*. 2015;abs/1508.07909. Available from: <http://arxiv.org/abs/1508.07909>.
- [42] AI S. Comparing BERT and GPT-2 as Language Models to Score the Grammatical Correctness of a Sentence; 2020. <https://www.scribendi.ai/comparing-bert-and-gpt-2-as-language-models-to-score-the-grammatical-correctne>
- [43] Irene Solaiman MB Jack Clark. GPT-2: 1.5B Release; 2019. <https://openai.com/blog/gpt-2-1-5b-release/>.
- [44] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*. 2018;abs/1810.04805. Available from: <http://arxiv.org/abs/1810.04805>.
- [45] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*. 2019;abs/1907.11692. Available from: <http://arxiv.org/abs/1907.11692>.

- [46] Murarka A, Radhakrishnan B, Ravichandran S. Classification of mental illnesses on social media using RoBERTa. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. online: Association for Computational Linguistics; 2021. p. 59-68. Available from: <https://aclanthology.org/2021.louhi-1.7>.
- [47] Pritzkau A. NLytics at CheckThat! 2021: Multi-class fake news detection of news articles and domain identification with RoBERTa-a baseline model. In: CLEF (Working Notes); 2021. p. 572-81.
- [48] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR. 2019;abs/1910.10683. Available from: <http://arxiv.org/abs/1910.10683>.
- [49] Shoenberger M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. CoRR. 2019;abs/1909.08053. Available from: <http://arxiv.org/abs/1909.08053>.
- [50] Mohammad Shoenberger RPPLJC Mostafa Patwary, Catanzaro B. State-of-the-Art Language Modeling Using Megatron on the NVIDIA A100 GPU; 2020. <https://developer.nvidia.com/blog/language-modeling-using-megatron-a100-gpu/>.
- [51] Smith S, Patwary M, Norick B, LeGresley P, Rajbhandari S, Casper J, et al. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. CoRR. 2022;abs/2201.11990. Available from: <https://arxiv.org/abs/2201.11990>.
- [52] Wang W, Chen L, Thirunarayan K, Sheth AP. Harnessing Twitter "Big Data" for Automatic Emotion Identification. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing; 2012. p. 587-92.
- [53] Goffman E. The presentation of self in everyday life; 1959.
- [54] Bazarova NN, Choi Y, Schwanda Sosik V, Cosley D, Whitlock J. Social Sharing of Emotions on Facebook; 2015. p. 154-64.
- [55] Data Protection Act 2018; 2018. <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>.
- [56] Mikulincer M, Shaver PR. Attachment theory and emotions in close relationships: Exploring the attachment-related dynamics of emotional reactions to relational events. Personal Relationships. 2005;12(2):149-68. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1350-4126.2005.00108.x>.
- [57] Dana Alon JK. GoEmotions: A Dataset for Fine-Grained Emotion Classification;. <https://ai.googleblog.com/2021/10/goemotions-dataset-for-fine-grained.html>.

- [58] Ramachandran VS, Jalal B. The Evolutionary Psychology of Envy and Jealousy. *Frontiers in Psychology*. 2017 Sep;8:1619. Available from: <http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01619/full>.
- [59] Beddiar DR, Jahan MS, Oussalah M. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*. 2021;24:100153. Available from: <https://www.sciencedirect.com/science/article/pii/S2468696421000355>.
- [60] Ng N, Yee K, Baevski A, Ott M, Auli M, Edunov S. Facebook FAIR's WMT19 News Translation Task Submission. *CoRR*. 2019;abs/1907.06616. Available from: <http://arxiv.org/abs/1907.06616>.
- [61] University P. "About WordNet."; 2010. <https://wordnet.princeton.edu/>.
- [62] Wang C, Cho K, Gu J. Neural Machine Translation with Byte-Level Subwords. *CoRR*. 2019;abs/1909.03341. Available from: <http://arxiv.org/abs/1909.03341>.
- [63] Kudo T, Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 66-71. Available from: <https://aclanthology.org/D18-2012>.
- [64] Tensorflow. Subword tokenizers, Tensorflow Guide;. https://www.tensorflow.org/text/guide/subwords_tokenizer.
- [65] Liu C, Zhang M, Fu Z, Hou P, Li Y. FLiText: A Faster and Lighter Semi-Supervised Text Classification with Convolution Networks. *CoRR*. 2021;abs/2110.11869. Available from: <https://arxiv.org/abs/2110.11869>.
- [66] Du J, Grave E, Gunel B, Chaudhary V, Celebi O, Auli M, et al. Self-training Improves Pre-training for Natural Language Understanding. *CoRR*. 2020;abs/2010.02194. Available from: <https://arxiv.org/abs/2010.02194>.
- [67] Mokiy V. Training Generalists in Higher Education: Its Theoretical Basis and Prospects. *Informing Science*. 2019 09;22:55-72.

Appendix A

Self-Attachment Technique Protocols

The following section outlines the 20 protocols of the self-attachment technique, which are integrated in the existing platform and recommended to the user to practice.

1. Connecting with the Child

Try to imagine the happy childhood photo/avatar and reflect on relevant positive affects, then imagine the unhappy photo and relevant negative affects. Repeat many times until this is easy to do. Try to imagine that the child, as you were, is near you (either in happy or unhappy state), and then imagine that you are embracing/cuddling the child. You can also imagine playing with the child.

2. Laughing at our Two Childhood Pictures

Begin by laughing at the childhood pictures, then think about why we laugh at these pictures, i.e. laugh at the contrast between them (Incongruity theory), laugh since your self is now superior than in the past (Superiority theory) and laugh because “Life is a tragedy when seen in close-up, but a comedy in long-shot” (Charlie Chaplin). Remember that we do not laugh at them to ridicule. This process will allow us to teach our childhood self to laugh.

3. Falling in Love with the Child

While looking at the happy childhood photo, recite selected happy love songs and imagine that you are establishing a deep emotional bond with the childhood self. Then sing with a loud voice, gradually using your whole body as if dancing with the child and having a loving dialogue.

4. Vow to Adopt the Child as Your Own Child

You imaginatively adopt your childhood self as your own child, loudly pledging to consistently support your child in any way possible. The pledge must be life-long and must be reinforced over time through practicing Self-Attachment protocols.

5. Maintaining a Loving Relationship with the Child

Choose a short phrase e.g. “You are my beloved” and repeatedly utter it while focusing on the happy and unhappy childhood photos. Recite one or two happy love songs, loudly repeating these using your whole body.

6. An exercise to Process the Painful Childhood Events

With closed eyes, recall a painful scene from childhood e.g. emotional or physical abuse in as much detail as possible, and associate the face of the child you were with your unhappy photo. After recalling this event and the related emotions, imagine your adult self approaching and embracing the child like a parent embracing a child in distress. While your eyes are still closed, continue to imagine supporting and cuddling the child, loudly supporting them (Examples: “Why are you hitting my child?” and “My darling, I will not let them hurt you more.”). Massage your face while doing so, which we interpret as cuddling the child.

7. Protocols for Creating Zest for Life

Using a mirror, imagine the reflection is your childhood self and loudly recite to it your selected happy love songs, using your entire body. Repeat songs and poems in many different circumstances e.g. while walking on the street or doing housework, to be able to integrate them into your life.

8. Loosening Facial and Body Muscles

You should loosen your muscles at least twice a day as you sing with your animated face and entire body, as if playing, dancing, laughing and having fun with the child as parents interact with children.

9. Protocols for Attachment and Love of Nature

To create an attachment with nature, you should visit a park or forest and spend time admiring nature, e.g. admiring a beautiful tree, as if seeing its branches and leaves at a deeper level for the first time. Repeat continuously and with different trees until you feel you have formed an attachment with nature. This will help to modulate your emotions and you will want to spend more time with nature each day.

10. Laughing at, and with One’s Self

Begin laughing with yourself about a small accomplishment e.g. in sports, housework, or any other task, however small or unimportant. With every small accomplishment, you should smile as if victorious, and gradually change this smile to laughter, and make this laughter last longer and longer. By practicing this you will be able to smile and laugh without ridicule about anything you have said or done in the past while maintaining compassion for your childhood self.

11. Processing Current Negative Emotions

With closed eyes, imagine the unhappy photo and project the unhappy emotions, e.g. anger, sorrow, towards the photo that represents the child. As with Type 6, we make contact with our adult self to attend to and care for the child to support the child and modulate the child's negative emotions. While projecting these negative emotions, loudly reassure the child and massage your own face, which we interpret as cuddling the child. Continue this until you have contained the negative emotions, at which point you can switch to focusing on the happy photo.

12. Continuous Laughter

At a time when you are alone, open your mouth slightly, loosen your face muscles, form a Duchenne smile and slowly repeat one of the following phrases as if laughing: eh, eh, eh, eh; ah, ah, ah, ah; oh, oh, oh, oh; uh, uh, uh, uh; or ye, ye, ye, ye. If a subject is needed for laughter, you can think about the silliness of the exercise. This exercise is a good antidote for stress.

13. Changing Our Perspective for Getting Over Negative Emotions

To break free of powerful negative patterns that emerge when we are stuck in the swamp of negative emotions, or a "psychological abyss", stare at the black vase in the Gestalt vase picture (below). When your perception changes and you see the white faces, convince yourself that these abysses can be overcome and try to laugh out loud as a victory sign. Having created a positive powerful pattern of love with the child through previous exercises, you can now depart from the field of negative patterns by singing your happy love song to enter the attractive field of love for the child instead. This is like changing our interpretation of the above image and instead of seeing a black vase of negative emotions discovering two white faces, you see the child and the adult self who are now looking at each other. Picture of the Gestalt vase:



[67]

14. Protocols for Socializing the Child

By repeating protocols 1-13 you can reduce negative emotions and increase positive affects. You are gradually able to perform these exercises with eyes open and can integrate them into your daily life. You should be able to extend compassion for the child to other people. The adult self should become aware of any narcissistic

tendencies or anti-social feelings of the child e.g. envy, jealousy, greed, hatred, mistrust, malevolence, controlling behavior and revengefulness. The adult self can behave like a parent to contain these emotions and discourage acting out any anti-social feelings and attitudes of the child by expressing affection to the child and simulating cuddles by massaging your face. The adult self should try to direct the child's anger and negative energy towards playing, creativity and development. As the child's positive affects increase and his/her negative affects decrease, by expressing positive emotions he/she can attract more positive reactions from others, and in turn gain a more positive outlook toward others.

15. Recognising and Controlling Narcissism and the Internal Persecutor

The adult self becomes aware of the facets of the trauma triangle: internal persecutor, victim, and rescuer. The adult self examines the effects of the triangle (narcissism, lack of creativity) in daily life and previous experiences. Your adult self can then review an important life experience and your social and political views as an adult, with awareness of how the internal persecutor operates. Your adult self can then create a list of examples from your experiences on how the internal persecutor operates, and carefully analyse these for examples of being drawn to trauma, being traumatized by the internal persecutor, and projecting the internal persecutor. You should be able to then reevaluate your own experiences, contain the internal persecutor and narcissistic tendencies and be able to develop creativity.

16. Creating an Optimal Inner Model

With awareness of the internal persecutor, we will recognise emotions of the child that were learned from parents or through interactions with them. With the guidance of the adult self, who can transfer compassion for the child to others, the child will learn to avoid projecting the internal persecutor (which would lead to them becoming the victim or rescuer).

17. Solving Personal Crises

In the midst of a personal crisis, as you continue to practice the protocol for modulating negative affects and the protocol for laughter, ask your child the following:

- How can you see the crisis as a way of becoming stronger? (ha ha ha)
- How can you interpret the crisis as a way of reaching your high goal? (ha ha ha)
- Has the internal persecutor been projecting onto others again?

The adult self asks the following questions:

- What is the similarity between this crisis and those faced before? How is it similar to the family crisis experienced as a child? Aren't the other person's

positive attributes greater than his/her negative ones? How would a mature person interpret the crisis in comparison to my child? Can I see it from the perspective of someone else? Can I put myself in their place and understand their affects? Given my new inner working model can I find a way to calm the people involved in the crisis so we can find a better solution for it?

- If not, can I respectfully maintain my distance and end the fight?

18. Laughing at the Harmless Contradiction of Deep-Rooted Beliefs/Laughing at Trauma

(i):

Laughing at the harmless contradiction of deep-rooted beliefs: “To those human beings who are of any concern to me I wish suffering, desolation, sickness, ill-treatment, indignities—I wish that they should not remain unfamiliar with profound selfcontempt, the torture of self-mistrust, the wretchedness of the vanquished: I have no pity for them, because I wish them the only thing that can prove today whether one is worth anything or not—that one endures.” This is meaningful with, “What doesn’t kill me makes me stronger.” Nietzsche’s wish is funny and a harmless contradiction of our deep-rooted beliefs. As we read the quote above, we remember our past sufferings and begin to laugh out loud when we get to “...I wish suffering...”

Laughing at trauma: First, visualize a painful event that took place in the distant past that you have struggled with for a long time, and despite its painfulness try to see it has led to some positive impact, lesson, change, or decision. We start with a painful event that happened in the distant past, so that by now we have been able to adjust our negative affects toward it. After repeated daily exercises, once we have experienced the forceful effectiveness of laughing at distant problems, we can gradually begin to laugh at more recent painful memories.

(ii):

Laughing at trauma: In expectation of hearing a funny joke we loosen our facial muscles, slightly open our mouths, and to grasp the incongruity in the joke we move our eyebrows up as a sign of surprise. As we repeat the sentences out loud, we slowly begin to laugh as we wait for the second part. And once we get to the first sentence of the second part, which is in complete contrast to our beliefs, we laugh out loud.

Not only should you: bear it, accept it, try to deal with it, tolerate its memory, try harder to endure its memory, adapt yourself to its memory, analyze and understand it and by doing so modulate your negative emotions and learn lessons for the future, try to soften your thoughts, depressive emotions, and anxieties, try to ... Like Nietzsche’s wish consider it a great treasure (ha ha ha...), cherish it with great love (ha ha ha...), welcome its challenges with all your heart (ha ha ha...), consider it a good omen with all your heart (ha ha ha...), consider its challenges a great fortune (ha ha ha...), celebrate its memory (ha ha ha...), celebrate its memory with great joy (ha ha ha...), consider it a true love (ha ha ha...), consider it a true love with great passion and intimacy (ha ha ha...) ... After repeated practice of the laughing exercises you can begin to apply it to things that worry you in the present and the future.

19. Changing Ideological Frameworks for Creativity

We challenge our usual ideological framework to weaken one-sided patterns and encourage spontaneity and examination of topics from multiple perspectives. Practice with subjects that you have deep-rooted beliefs and are excited about e.g. anything from political/social issues to ideas on marriage and sexuality. For instance, examine the topic of racism and consider whether you have any latent racism and consider this subject in the dual role of proponent and opponent. Repeat with topics where you may have stronger views e.g. marriage and sexual orientation. If you are politically in the center, consider the subject both from a leftist and rightist point of view and try to understand both sides of the issue and see the subject from three perspectives.

20. Affirmations

Put together a list of affirmations by different important figures. Choose ones that have an impact on you from the start and can provide you with strength in the long path for reaching your ultimate goal. Read them out loud. A few examples:

- “My formula for greatness in a human being is Amor Fati: that one wants nothing to be other than it is, not in the future, not in the past, not in all eternity.” (Nietzsche)
- “I assess the power of a will by how much resistance, pain, torture it endures and knows how to turn it to its advantage.” (Nietzsche)
- Life is not easy. At times we inevitably suffer from hopelessness and paranoia unless if we have an ideal goal that helps us surpass suffering, weakness, and betrayals.” (Bronstein)

Appendix B

Analysis of Emotion Datasets in Research Literature

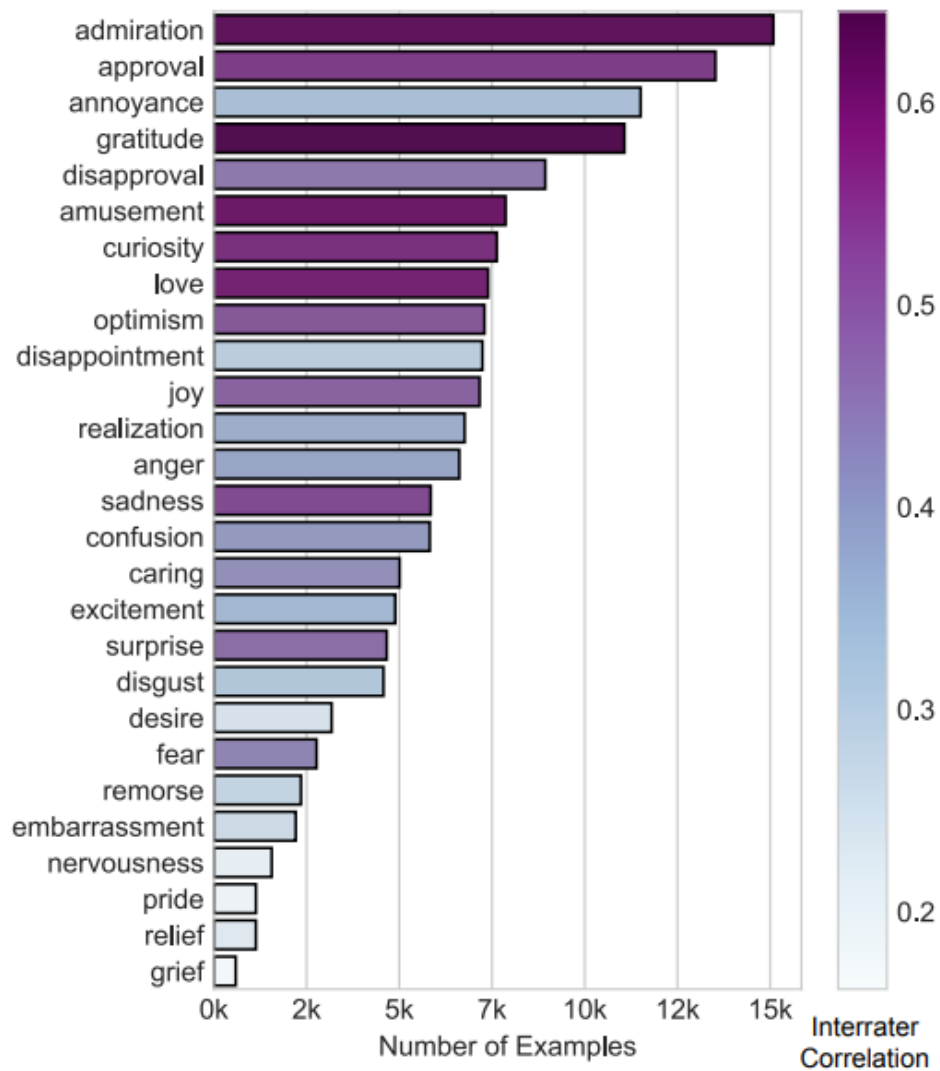


Figure B.1: Distribution of emotions in the GOEMOTIONS dataset based on number of examples and showing the interrater correlation by annotators [19].

Emotion	Precision	Recall	F1
admiration	0.53	0.83	0.65
amusement	0.70	0.94	0.80
anger	0.36	0.66	0.47
annoyance	0.24	0.63	0.34
approval	0.26	0.57	0.36
caring	0.30	0.56	0.39
confusion	0.24	0.76	0.37
curiosity	0.40	0.84	0.54
desire	0.43	0.59	0.49
disappointment	0.19	0.52	0.28
disapproval	0.29	0.61	0.39
disgust	0.34	0.66	0.45
embarrassment	0.39	0.49	0.43
excitement	0.26	0.52	0.34
fear	0.46	0.85	0.60
gratitude	0.79	0.95	0.86
grief	0.00	0.00	0.00
joy	0.39	0.73	0.51
love	0.68	0.92	0.78
nervousness	0.28	0.48	0.35
neutral	0.56	0.84	0.68
optimism	0.41	0.69	0.51
pride	0.67	0.25	0.36
realization	0.16	0.29	0.21
relief	0.50	0.09	0.15
remorse	0.53	0.88	0.66
sadness	0.38	0.71	0.49
surprise	0.40	0.66	0.50
macro-average	0.40	0.63	0.46
std	0.18	0.24	0.19

Figure B.2: Results from the experiments on fine-tuning a pre-trained BERT model with GOEMOTIONS [19].

Emotion	Most-used speaker words	Most-used listener words	Training set emotion distrib
Surprised	got,shocked,really	that's,good,nice	5.1%
Excited	going,wait,i'm	that's,fun,like	3.8%
Angry	mad,someone,got	oh,would,that's	3.6%
Proud	got,happy,really	that's,great,good	3.5%
Sad	really,away,get	sorry,oh,hear	3.4%
Annoyed	get,work,really	that's,oh,get	3.4%
Grateful	really,thankful,i'm	that's,good,nice	3.3%
Lonely	alone,friends,i'm	i'm,sorry,that's	3.3%
Afraid	scared,i'm,night	oh,scary,that's	3.2%
Terrified	scared,night,i'm	oh,that's,would	3.2%
Guilty	bad,feel,felt	oh,that's,feel	3.2%
Impressed	really,good,got	that's,good,like	3.2%
Disgusted	gross,really,saw	oh,that's,would	3.2%
Hopeful	i'm,get,really	hope,good,that's	3.2%
Confident	going,i'm,really	good,that's,great	3.2%
Furious	mad,car,someone	oh,that's,get	3.1%
Anxious	i'm,nervous,going	oh,good,hope	3.1%
Anticipating	wait,i'm,going	sounds,good,hope	3.1%
Joyful	happy,got,i'm	that's,good,great	3.1%
Nostalgic	old,back,really	good,like,time	3.1%
Disappointed	get,really,work	oh,that's,sorry	3.1%
Prepared	ready,i'm,going	good,that's,like	3%
Jealous	friend,got,get	get,that's,oh	3%
Content	i'm,life,happy	good,that's,great	2.9%
Devastated	got,really,sad	sorry,oh,hear	2.9%
Embarrassed	day,work,got	oh,that's,i'm	2.9%
Caring	care,really,taking	that's,good,nice	2.7%
Sentimental	old,really,time	that's,oh,like	2.7%
Trusting	friend,trust,know	good,that's,like	2.6%
Ashamed	feel,bad,felt	oh,that's,i'm	2.5%
Apprehensive	i'm,nervous,really	oh,good,well	2.4%
Faithful	i'm,would,years	good,that's,like	1.9%

Figure B.3: Distribution of emotions in the EMPATHETICDIALOGUES dataset based on their representation and showing the most frequently appearing speaker and listener words [20].

Model	Candidate Source	Retrieval		Retrieval w/ BERT		Generative	
		P@1,100	AVG BLEU	P@1,100	AVG BLEU	PPL	AVG BLEU
Pretrained	R	-	4.10	-	4.26	27.96	5.01
	ED	43.25	5.51	49.94	5.97	-	-
Fine-Tuned	ED	56.90	5.88	65.92	6.21	21.24	6.27
	ED+DD	-	5.61	-	-	-	-
	ED+DD+R	-	4.74	-	-	-	-
EmoPrepend-1	ED	56.31	5.93	66.04	6.20	24.30	4.36
TopicPrepend-1	ED	56.38	6.00	65.96	6.18	25.40	4.17

Figure B.4: Evaluation results from several configurations run with the EMPATHETICDIALOGUES dataset based on a 4-layer custom transformer architecture [20]. Metrics include precision retrieving the correct test candidate out of 100 test candidates (P@1,100), average of BLEU-1,-2,-3,-4 (AVG BLEU) and perplexity (PPL).

Appendix C

Conversation Flow

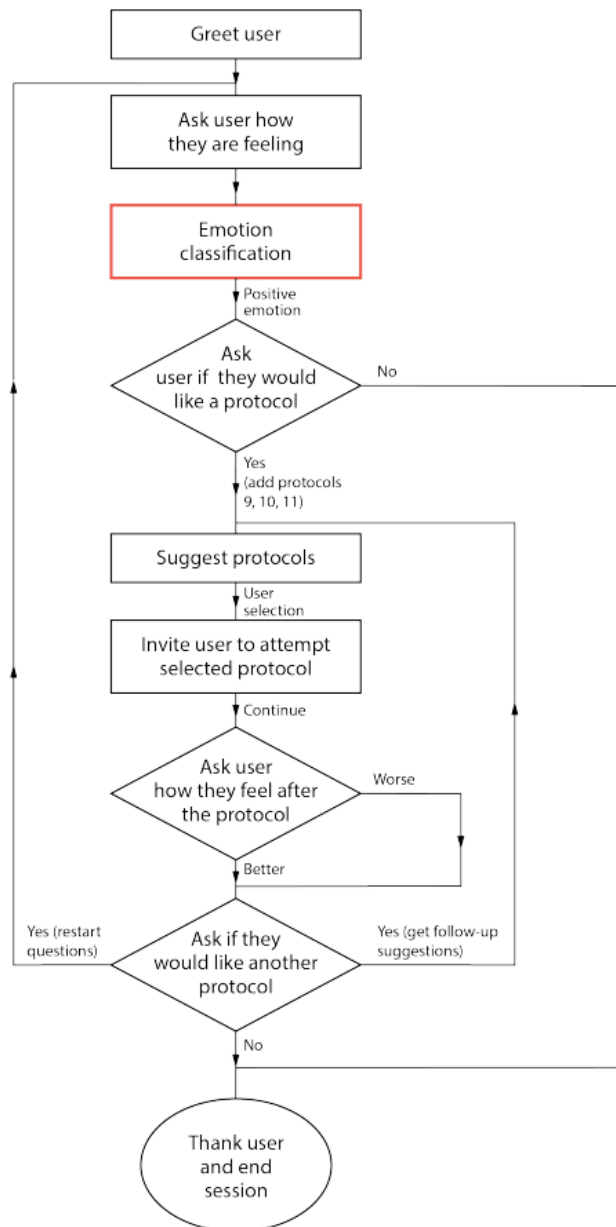


Figure C.1: Graph of the conversation flow in a positive emotional context [15].

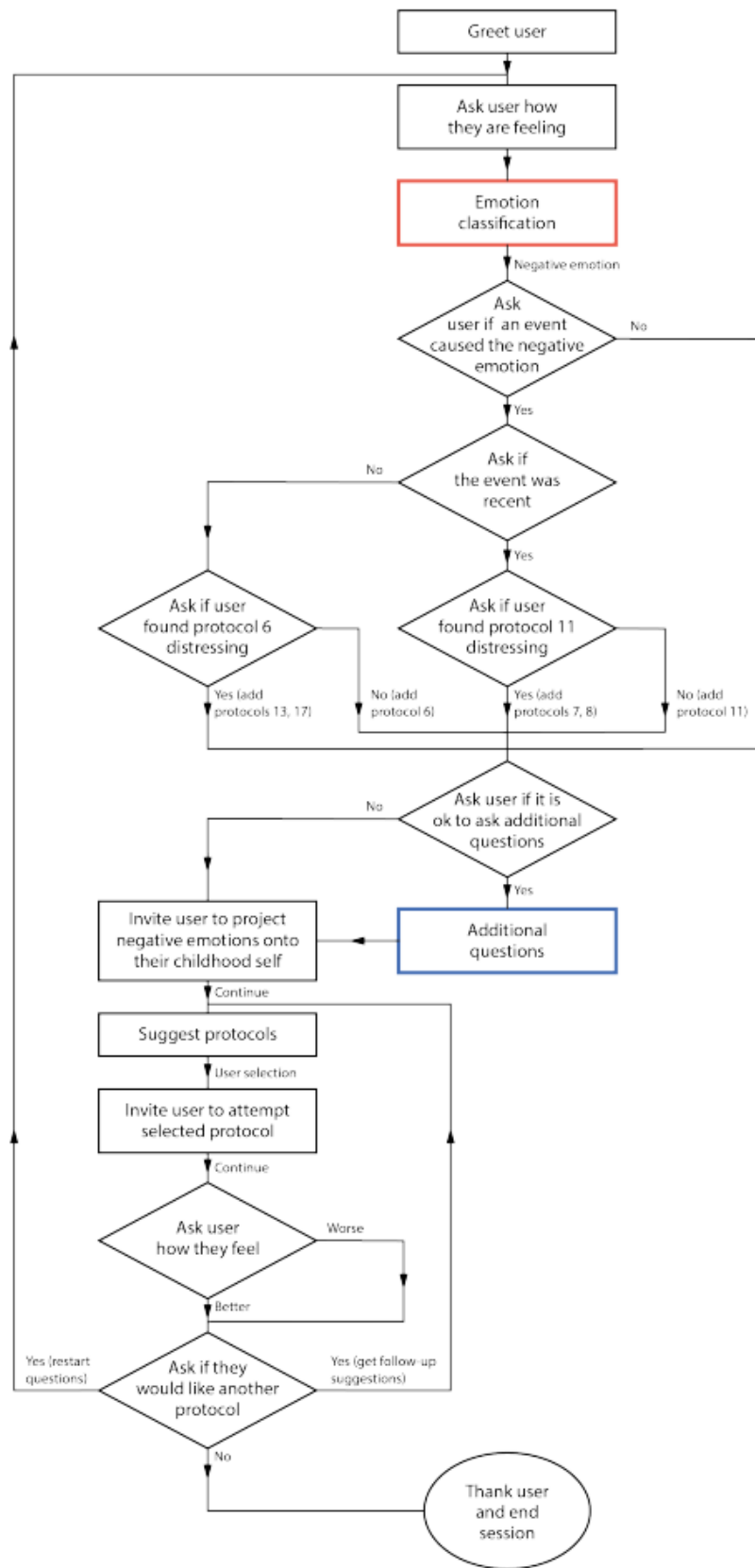


Figure C.2: Graph of the conversation flow in a negative emotional context [15].

Appendix D

Example Survey Entries from Data Crowd-Sourcing

Q100. Writing User Expressions

Please write 3 different user prompts that would follow the question "How are you feeling today?" These prompts should consider different situations and you should respond as if you are disgusted.

Prompt 1	I feel very uncomfortable to how my best friend has responded in a situation.
Prompt 2	I find it very difficult to accept the way my best friend think about what happened yesterday.
Prompt 3	I am disgusted by the way people think about this issue!

Figure D.1: Example of an accepted and quality response from the crowd-sourcing of disgust and disappointment.

Q100. Writing User Expressions

Please write 3 different user prompts that would follow the question "How are you feeling today?" These prompts should consider different situations and you should respond as if you are disgusted.

Prompt 1	I feel sad because I cant find a job.
Prompt 2	I feel like i'm stressed all the time because of my job.
Prompt 3	Sad, i dont know if i should stay with my boyfriend.

Figure D.2: Example of low quality response from the crowd-sourcing of disgust and disappointment. This entry was accepted based on the rejection criteria but the prompts for the emotion were deleted and considered as empty as they did not represent the emotional context necessary.

Q100. Writing User Expressions

Please write 3 different user prompts that would follow the question "How are you feeling today?" These prompts should consider different situations and you should respond as if you are disgusted.

Prompt 1	Im ok
Prompt 2	Have been better
Prompt 3	A little bit sad

Figure D.3: Example of a rejected low quality response from the crowd-sourcing of disgust and disappointment.