Imperial College London

Department of Computing

A deep-learning assisted empathetic guide for self-attachment therapy

Lisa Alazraki

Supervisor:
Professor Abbas Edalat

Second marker:
Dr Anandha Gopalan

Submitted in partial fulfilment of the requirements for the MSc Degree in
Computing (Artificial Intelligence and Machine Learning)
of
Imperial College London

September 2021

Abstract

Conversational agents are increasingly being used to administer psychotherapy, both within academia and commercially. At present, the majority of chatbots built for this purpose are rule-based, as this framework guarantees the stability and predictability that are required in a sensitive area such as mental healthcare.

In this work, we present a new dataset and a computational strategy that aim to improve users' experience of interacting with a virtual agent for guiding them through self-attachment therapy, while at the same time maintaining safety and reliability. Our framework combines a deep-learning classifier for identifying the underlying emotion in a user's text response, as well as a deep-learning assisted retrieval method for producing novel, fluent and empathetic utterances.

We also craft a set of human-like personas for our chatbot. Our goal is to achieve a high level of user engagement during virtual therapy sessions.

We evaluate the effectiveness of our framework in a non-clinical trial with N=16 participants, all of whom have had at least four interactions with the bot over the course of five days. We find that our platform is consistently rated higher than the baseline for empathy, user engagement and usefulness.

Finally, we provide guidelines to further improve the design and performance of the chatbot, in accordance with the feedback received.

Acknowledgements

First and foremost, the completion of this project could have not been accomplished without the continuous support of my advisor, Professor Abbas Edalat. To him goes my deepest gratitude.

I would also like to extend my thanks to Dr Anandha Gopalan, for his precious feedback on my background and progress report and for pointing me in the direction of relevant literature.

Moreover, I would like to express my sincere gratitude to Professor Foaad Khosmood of California Polytechnic State University for the generosity and patience with which he has shared with me his expertise in natural language processing.

Besides the professors that have guided me through this academic journey, I am thankful to Ali Ghachem, who created the first SAT chatbot and has been incredibly helpful in explaining to me every detail of his work, Neophytos Polydorou, who has had the patience to read my drafts and provided very valuable feedback, and Adrian Millea, for his useful feedback and for answering many of my technical questions in the early stages of the project.

My gratitude also extends to Noor Research (Atlanta, USA), for generously funding the collection of the `EmpatheticPersonas` dataset.

Finally, I would like to thank DeepMind for allowing me the opportunity to complete my postgraduate degree through financial support and mentorship. A very special thanks goes to my mentor at DeepMind, Dr Mark Rowland, who has provided me with invaluable advice and whose continuous encouragement throughout this atypical and challenging academic year has allowed me to stay focussed and motivated, ultimately leading to the successful completion of this project.

Contents

1 Introduction

It is estimated that almost a billion people worldwide – approximately 13 percent of the global population – suffer from at least one mental disorder (1). This number has increased by a third since 1990, and it is expected to continue to grow at an even steeper rate in the near future, due to the direct and indirect effects of the COVID-19 pandemic (2; 3). Research has shown that mental illness is far from being evenly distributed across socio-economic strata, as individuals from low-income households are more likely to be exposed to risk factors such as domestic violence (4), worker exploitation (5), incarceration (6), deprivation and homelessness. The World Health Organization estimates that, as a result of this exposure, individuals from low socio-economic groups are significantly more likely to experience depression and other mental health conditions (7).

Despite the demonstrated need for pervasive, affordable mental healthcare, the considerable personal financial cost that is often associated with traditional psychotherapy prevents patients from low-income backgrounds from accessing the specialised care that they require (8; 9). In addition, the ratio of mental health professionals per 100,000 population in low and middle-income countries is estimated to be only 0.5 percent of that of high-income countries, resulting in a further barrier to accessing therapy in many areas of the world (10).

Confronted with these issues, researchers have looked at digital technology as a means to deliver mental health services to the wider population (11; 12). While technology alone may not be an adequate substitute for the beneficial relationship that forms between a patient and their therapist (13; 14), it remains a promising tool for augmenting existing approaches, due to the broad availability of the internet and the extensive global distribution of digital devices (it is estimated that over 80 percent of the population in low and middle-income countries owns a mobile phone) (15). As a result, a wide range of technological tools aimed at mental health support have been investigated and deployed within academia and industry (16; 17), many of which take the form of conversational agents administering various forms of psychotherapy (18).

It should be noted, however, that using conversational agents in a sensitive area such as mental healthcare poses significant challenges. Current deep-learning approaches to text and speech generation lack the necessary oversight to prevent a system from producing output that is insensitive (19) and even offensive (20), and thus potentially damaging to a patient's well being. A recent literature review study has observed that the large majority of mental-health-oriented chatbots currently in existence do not use machine learning at all, favouring more stable and predictable techniques such as rule-based modelling (21). On the other hand, purely rule-based bots have a limited, keyword-based understanding of user input and their dialogue can be perceived as simplistic, monotonous and devoid of empathy, resulting in a failure to fully engage users (22; 23).

In this paper, we present a computational framework for the delivery of self-attachment technique (SAT), a recently developed form of psychotherapy. Our approach is aimed at maintaining the safety of rule-based strategies while also ensuring that the conversational agent generates responses that are empathetic, diverse and fluent, as well as appropriate to the user's emotional state. To this end, we create a new dataset – `EmpatheticPersonas` – of 1,181 verbal expressions of emotion and 2,143 empathetic rewritings of base utterances, both crowd-sourced. We devise a strategy for generating, at each point in the conversation flowchart, novel yet safe utterances, trying to minimise any unpredictability in their overall meaning. To do so, we extract short, self-contained sentences from the set of utterance rewritings in the `EmpatheticPersonas` dataset, by splitting each of them at major punctuation marks (full stops, question marks and exclamation points). We then join together the extracted sentences in all possible sequential combinations and obtain a large corpus of new utterances. From this corpus, the conversational agent retrieves – through a multi-objective function that simultaneously maximises empathy, fluency and novelty – the most appropriate utterance to present to the user. To compute the empathy score of an utterance, we use a T5 model (24) fine-tuned on a labelled subset of our dataset (~80% acc., ~81% macro-f1); for the fluency score we add a penalty for each repeated word within an utterance to its perplexity value generated by a GPT-2 language model (25); finally, to obtain the novelty score, we compute the overlap distance over all possible n-grams between an utterance and each of the agent's previous utterances.

In addition, we adopt a RoBERTa model (26) for the task of emotion recognition (~95% acc., ~95% macro-f1), training it on the existing Emotion dataset (27) and further fine-tuning it on the expressions of emotion in our own corpus. This allows the bot to identify a user's emotional state from their text responses and answer accordingly.

Lastly, we craft human-like characters for our conversational agent which users can choose from and interact with. We evaluate the application through a human trial with N=16 subjects from the non-clinical population, as well as a medical professional specialised in mental health.

We show that our approach is scored highly for perceived usefulness, ability to communicate empathetically and user engagement, and that it performs significantly better than a previous rule-based version of the SAT chatbot (28) in all three areas. Our agent's ability to recognise human feelings is also assessed positively, with 63% of trial participants agreeing that the bot was successful in guessing their emotions.

In light of the feedback received during the trial, we conclude with a reflection on the strengths of our work as well as the weaknesses, and we draw a list of possible changes, improvements and extensions which we believe may benefit the chatbot and its users.

2 Background

The work presented in this paper builds upon existing research in natural language processing relating to chatbot-assisted psychotherapy. In particular, we extend and enhance a previous, strictly rule-based version of the SAT chatbot.

2.1 Existing approaches to chatbot-assisted mental support

Many of the mental health support chatbots currently in existence approach dialogue generation using a tree-structured flowchart, whose transitions between prearranged states are determined by user input (29; 30; 31; 32; 33; 34; 35; 36). This input can take the form of open text (33; 34), selection from a predetermined set of options (29), or a combination of the two (30; 31; 32).

Within this framework, the conversation can be modelled as a slot-filling problem, where the user's input is integrated into pre-existing templates to create a chatbot utterance (33; 36; 34). Alternatively, it can be informed by completely fixed, predetermined utterances, often written by mental health professionals with formal psychology training (30; 31). Using fixed templates and utterances enables researchers to maintain control over the dialogue, ensuring that the bot will not deliver insensitive or problematic responses which could potentially have a negative effect on the patient's mental health. However, this can also render the experience less engaging due to the conversation appearing rigid and repetitive, especially if a user interacts with the chatbot on a regular basis (37). To introduce a degree of variety in the conversation, Ghandeharioun et al. propose a retrieval method that randomly selects each bot utterance from a set of variations (35); however, the set only comprises six options, and thus it is unlikely to be able to prevent the dialogue from becoming repetitive over time.

2.2 Digital psychotherapy and self-attachment technique

Self-attachment technique (SAT) is a recently developed psychotherapy framework consisting of self-administered protocols aimed at establishing and reinforcing neural patterns associated with secure attachment (38). It stems from findings in developmental psychology that link insecure attachment in childhood with affective disorders in adulthood. In SAT, the patient simultaneously enacts the roles of the adult and the childhood self, gradually building a bond between the two. This allows the childhood self to become securely attached to the adult self, enhancing positive emotions and equipping the patient with the cognitive tools to better tackle challenging situations and negative feelings. Appendix A contains a brief overview of the SAT protocols.

SAT is suitable to be dispensed in a digital, automated manner due to its self-administered nature (39). A recent study investigating the applicability of a chatbot for the delivery of SAT protocols received some encouraging results, with the majority of survey respondents rating the platform as useful (28). On the other hand, the entirely rule-based bot was deemed to be empathetic by only 29 percent of respondents, and only 28 percent agreed that conversing with it was an engaging experience. Here, we extend the previous work done on the SAT chatbot by leveraging deep learning methods for emotion recognition and utterance retrieval. Our goal is to increase user's perception of empathy and overall engagement.

2.3 The importance of empathy

According to psychotherapy research, the most important factor to ensure the establishment of a beneficial relationship between a therapist and their patient is the ability of the former to engage in an empathetic manner with the latter (40). It should be noted that empathy is a heterogeneous concept that has been given multiple definitions across literature and is comprised of different aspects and elements (41; 42). Here we refer in particular to the work of Godfrey T. Barrett-Lennard, who identifies three main phases of an empathetic dialogue between two individuals: a first phase where the listener sympathises and resonates with what is being expressed by the speaker, a second phase in which the listener compassionately responds to the speaker, and finally a third phase where the speaker assimilates the listener's response (43). Similarly to other researchers who tackled the problem of empathy in conversational agents before us (44), we will mainly focus on Barrett-Lennard's second phase – the *expressive* phase of empathetic exchange – in an attempt to create a virtual psychotherapist able to display compassion toward the user.

To this end, we hire crowd-workers on popular micro-task platforms to rewrite a set of base utterances in an empathetic manner. We give minimal, yet specific instructions and guidance on the emotional state of the interlocutor to whom those utterances would be directed. These rewritings, collected in the `EmpatheticPersonas` dataset, form the basis from which the chatbot's empathetic responses are generated.

2.4 Virtual personas and user engagement

According to past research, users' experience of interacting with a chatbot improves significantly when this is equipped with a coherent identity (45). Moreover, psychology studies have highlighted that individuals tend to prefer psychotherapists of a certain age or sex according to several factors. For example, women generally report higher levels of comfort when self-disclosing to female practitioners compared to male ones (46), and patients tend to choose younger or older specialists depending on the particular issue that they are facing (older therapists are preferred for universal problems such as mourning, while younger ones are favoured when dealing with issues that more typically affect young people, such as heartbreak or cyberbullying) (47).

Having collected the age range and sex of each crowd-worker who contributed to the `EmpatheticPersonas` dataset, we craft different personas for our conversational agent. We create a younger female persona named Olivia, whose dialogue is based on the empathetic rewritings provided by female crowd-workers aged 18 to 39, a younger male persona named Arman, whose utterances are generated from those of the male respondents in the same age range, an older female persona named

Gabrielle, whose dialogue is based on the rewritings by female survey respondents aged 40 to 69, and finally an older male persona named Robert, whose interactions are crafted from the survey responses given by male crowd-workers aged between 40 and 69. Users are able to select the persona that they wish to speak to at the beginning of each session, according to their mood and preference.

Finally, we also create a more generic, less human-like identity for our chatbot, Kai, whose dialogue is conditioned on the whole dataset and is not associated to any sex or age range. We ask trial participants to evaluate Kai and the other personas separately, in order to assess whether associating human-like characters to the chatbot can improve the level of user engagement and the perceived empathy of an interaction.

2.5 Privacy and ethics

Like many applications of modern computing, virtual psychotherapy is linked to potential legal, social and ethical issues, further amplified by the sensitive nature of mental health treatment. Some of the confidential topics that a patient may discuss with the chatbot qualify as personal data under the UK Data Protection Act (DPA) 2018 (48) and the UK General Data Protection Regulation (GDPR) (49). In addition, the responses dispensed by an automated system can be unpredictable, clumsy and even harmful, and in themselves reason for ethical concern.

In particular, our bot is not equipped to assist patients who suffer from depression or other serious mental health conditions, nor has it been tested on the clinical population. Its responses may be inappropriate and potentially damaging within contexts involving violence or self-harm. A careful and considered approach should be taken when dealing with users that may be experiencing mental distress, and future research should meticulously assess any risks associated with using the platform in a clinical setting as well as the appropriate solutions, preferably in collaboration with trained mental health professionals.

Finally, it is worth stressing that, during interactions, the platform does not store any user input, personal information or insight such as geolocation data, IP or MAC addresses, IMEI codes or any other type of metadata from users' devices.

Ethical approval

The collection of the `EmpatheticPersonas` dataset was approved by the Research Ethics Committee of Imperial College London. In accordance with data protection laws, the crowd-worker platform IDs that were automatically obtained during the surveys were deemed to be personally identifiable information and thus not included in the dataset. All crowd-workers have authorised the handling of their responses.

The non-clinical trial for the evaluation of the SAT chatbot also received ethical approval from Imperial College's Research Ethics Committee. In compliance with current regulations, the information collected during the trial is limited to the minimum required for the successful completion of the study. Moreover, every participant has been advised on how this information will be handled before giving consent, and is able to access it at all times or delete it from the system.

3 Dataset and data collection

3.1 Survey preparation

We prepared four surveys for crowd-sourcing the `EmpatheticPersonas` dataset – two relating to the emotional contexts of sadness and anger, and two for the contexts of anxiety/fear and happiness/content[1]. Each survey contains two tasks: one asking the respondents to provide multiple textual expressions of emotion (answering the question 'How are you feeling?') for two different emotional contexts, and one requiring them to rewrite base utterances to render them empathetic, keeping in mind that such utterances would be directed to an interlocutor who is experiencing a specified emotion among the four listed above. In addition, we ask respondents to provide information about their sex and age group.

3.2 Recruitment of survey respondents

Survey respondents were recruited via the crowd-working websites Amazon Mechanical Turk (50) and Prolific (51). Responses were rejected if they amounted to less than 50 percent of the survey, if they contained poorly written syntax or unrelated text, or if the base questions that were meant to be rewritten had been copy-pasted without changes. In all the other cases, the responses were accepted. Where minor grammar, syntax or semantic mistakes were present, these were rectified before insertion into the dataset. In total, 200 survey responses were accepted – 50 for each of the four surveys.

3.3 Data analysis

The `EmpatheticPersonas` dataset comprises 200 rows, each corresponding to a survey response. Each row contains the sex and age range of that respondent, as well as the expressions of emotions and empathetic rewritings that they provided. There are two sexes (male, female) and six age groups within the corpus. While the distribution of data samples across the two sexes is balanced (98 females and 102 males), the majority of the samples originate from the 30-39 and 40-49 age groups for both sexes, as shown in Figure 1.

There are 1,181 textual expressions of emotion in the dataset, distributed across four emotional contexts – sadness, anger, fear/anxiety, joy/content – and 2,143 empathetic rewritings of 45 base utterances. Each subset of 50 rows collects the responses to one of the four surveys. It is worth noting that each survey contains different emotional contexts as well as different base utterances to rewrite, and thus so does each subset of rows. A complete summary of the contexts and utterances present in each row subset within the dataset, together with the number of data samples in each column, is visible in Appendix B.

[1] The first two surveys were written in collaboration with Imperial College students Ali Ghachem and Neophytos Polydorou.

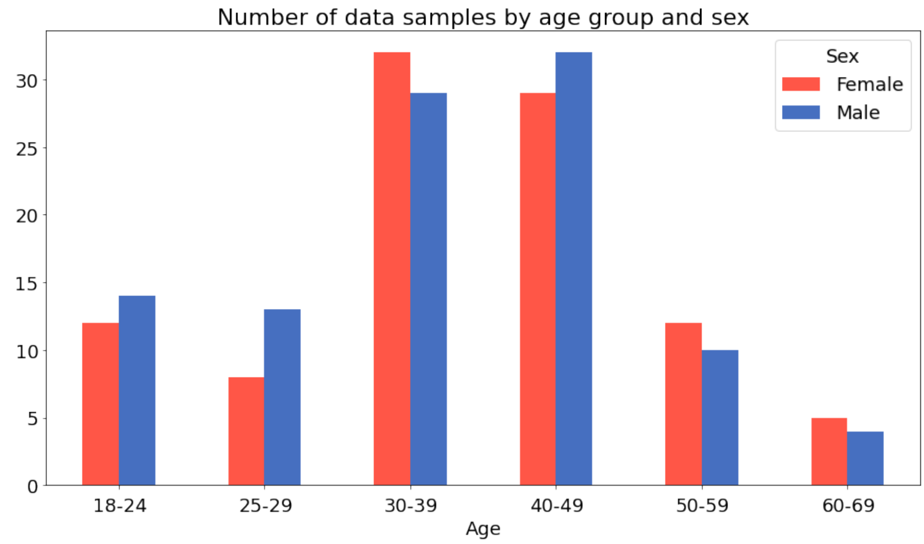

Figure 1: Age distribution for both sexes across samples in the `EmpatheticPersonas` dataset, showing that most samples belong to the middle age groups 30-39 and 40-49.

Accounting for some missing data, the corpus contains between 42 and 50 rewritings for each base utterance. It also contains, in total, 285 textual expressions of fear/anxiety, 297 expressions of anger, 299 expressions of sadness and 300 expressions of joy/content. All empty cells are filled with NaN values.

3.4 Empathy annotation

It should be noted that the utterance rewritings in our corpus may convey different levels of empathy. This is due to the individual personality of each survey respondent and their interpretation of the task, as well as the fact that we did not reject responses based on their perceived level of empathy. In order to build an effective empathy classifier, necessary to ensure that our system produces the most appropriate responses, we create a separate dataset by randomly extracting 1,100 utterance rewritings from our corpus and annotating them for empathy, using discrete numerical labels from 0 to 2 (where 0 corresponds to a non-empathetic utterance and 2 to a strongly empathetic one). We use this scale as it correlates with the existing EPITOME dataset (44), which is part of the empathy classifier training pipeline described in Subsection 5.4.

To avoid excessively biasing the model toward our own judgement, we also enlist two volunteer annotators to re-score the 1,100 rewritings for empathy, using the same scale. Both volunteers have worked in healthcare and are familiar communicating empathetically with patients. For each rewriting, we compute the overall empathy score by choosing the majority label out of the three individual ones. If all three labels are different, we assign a score of 1.

It should be noted that this labelling method may still invite bias, as all three annotators belong to similar age groups (30-39 and 40-49). In future implementations, it is recommended that the rewritings are re-scored for empathy via crowd-sourcing.

4 Tools

In this section, we provide some technical background for the natural language processing (NLP) tools used in our framework.

4.1 tf-idf

The statistical measure tf-idf is obtained by multiplying the term frequency (tf) of each word in a document (52) by its inverse document frequency (idf) within the entire corpus (53). In practice, the tf-idf of a word t in a document containing m words within a corpus containing M documents is obtained applying the formula

$$\text{tf-idf}(t, m, M) = \text{tf}(t, m) \times \log\left(\frac{M}{\text{df}(t)}\right),$$

where $\text{tf}(t, m)$ is the number of occurrences of that word within that document divided by the total number of words in the document, and $\text{df}(t)$ is the number of documents in the corpus containing that word.

The tf-idf measure assigns a numerical value to each word relating its importance within a document to its rarity across the corpus, thus allowing a text to be mapped to a numerical vector and be processed by a statistical model. In this work, we use tf-idf as a starting point to build a logistic regression classifier, which we will utilise as a baseline for our empathy classification experiments.

4.2 Language models and perplexity

In simple terms, a language model is a statistical tool that computes the probability of a sequence of words based on how common that sequence is in a training corpus (54). For example, given a sequence containing m words $t_1, ..., t_m$, we can use a language model to assign it a joint probability $P(t_1, ..., t_m)$. A language model also allows picking the most likely next word from a vocabulary V to extend this sequence, choosing the word t in V that gives the greatest conditional probability $P(t_{m+1} = t \,|\, t_1, ..., t_m)$.

To evaluate the performance of a trained language model, we apply it to a new corpus and measure the resulting perplexity (55). Perplexity is the inverse of probability, thus the higher the perplexity returned by a model, the less effective it is at predicting the new corpus and the worse its performance. Of course, this assumes that the corpus contains documents that are syntactically correct and whose perplexity value should indeed be low. On the other hand, if we are in doubt about the correctness of a document, we can use perplexity to assess it, provided that the value has been generated by an accurate enough language model.

Language models can be n-gram based (56) or neural (57) – the latter usually relying on a recurrent neural network (RNN) architecture (58; 59) or a transformer one (60; 61). Here we focus solely on transformer-based language models, pre-trained on large unlabelled text corpora. These models can then be fine-tuned for a classification task on a labelled dataset, achieving far better results than networks trained on the labelled corpus alone (62).

4.3 GPT-2

GPT-2 is a neural language model that uses a decoder-only transformer architecture (25). Its smallest version (and the one we will use in our implementation) has 12 layers and 117 million parameters (63) trained on the WebText dataset – a large unlabelled corpus of eight million web pages.

GPT-2 is a causal, unidirectional language model: when computing the probability of a word in a sequence, it only takes into account text that is to the left of that word. Causal language models are built for generating text that is correct and fluent; a recent comparison between GPT-2 and the masked language model BERT (64) showed that the former returns better results when tasked with scoring sentences from the CoNLL-2012 dataset (65) for syntactic correctness (66). In our framework, we use the perplexity generated by a GPT-2 language model to evaluate the fluency of a sentence.

4.4 RoBERTa

RoBERTa is a transformer language model that aims to optimise and improve BERT's architecture (26). While comparable to BERT in its encoder-only structure, it is trained for longer using much larger mini-batches and sequences, on a sizeable collection of unlabelled text datasets. It also introduces a dynamic word-masking strategy that changes every four training epochs (rather than occurring only once during preprocessing as is the case in BERT). Moreover, it does not attempt to predict whether different sequences belong to the same document, relinquishing the next sentence prediction (NSP) loss that is a staple of BERT.

Thanks to these improvements, RoBERTa beats BERT's performance in most benchmarks, establishing itself as a highly effective model for many NLP tasks. In our pipeline, we experiment with the RoBERTa base model, which has 12 layers and 125 million parameters (63), to classify underlying emotions in text as well as the level of empathy of an utterance.

4.5 T5

T5 is a neural language model with an encoder-decoder transformer architecture (24). This model takes a text string as input and returns a text string as output, unlike RoBERTa and other BERT-based models which can only return a single prediction and whose output is therefore constrained to either a class label or a specific span of the input. Its text-to-text structure gives T5 great flexibility, allowing it to be used with minimal changes on a variety of tasks such as machine translation, classification, document summarisation and even regression (provided numerical values are first converted to their string representation). According to its proponents, T5 achieves state-of-the-art performance in many benchmarks.

T5's base version comprises a 12-layer encoder and a 12-layer decoder. Its 220 million parameters (63) are trained on the open-source Colossal Clean Crawled Corpus (C4). We use this model to classify emotion and empathy in text, comparing its results with those obtained by RoBERTa.

5 Implementation

All the code described in this section is publicly available at `https://github.com/LisaAlaz/SATbot`. Appendix C contains two sample conversations between the chatbot and a human.

5.1 Protocol recommendation strategy

The main objective of the SAT chatbot is to recommend the most appropriate self-attachment technique protocols to its users. In order to do so, the bot must first understand each user's circumstances by asking questions about their emotional state and any events that may have caused it.

In this implementation, we make only minimal changes to the existing protocol recommendation mechanism designed by Ghachem for the earlier version of the application (28), as our framework is mainly focussed on the chatbot's dialogue. The mechanism dynamically draws a list of protocol suggestions, adding specific recommendations according to the answers provided by the user. In accordance with feedback given to the previous version of the application by evaluation trial participants, which indicated that the suggestions were at times too rigid and some SAT protocols were never recommended, we make changes to this strategy to add up to four additional protocols to the list of recommendations whenever this has less than four elements already in it. These additional suggestions are randomly drawn from a pool of protocols aimed at enhancing positive emotions, taking care that the final list contains no duplicate protocols.

5.2 Emotion recognition

In order to customise the dialogue to a user's emotional state, the SAT chatbot asks the question 'How are you feeling?' at the beginning of each conversation, immediately after the user has chosen the persona that they wish to interact with. Consistently with the data collected in the `EmpatheticPersonas` dataset, we aim to be able to discern between four emotional contexts: happiness/content, sadness, anxiety/fear and anger.

To achieve effective emotion recognition given a user's text response, we experiment with different models and compare their results. In accordance with existing research in emotion and sentiment analysis (27; 67; 68), we apply a light preprocessing to the text data (punctuation removal and lower-casing) before tokenization.

Baseline

We define as our baseline the rule-based emotion classifier implemented in the previous version of the SAT chatbot (28). This method defines sets of 'for' and 'against' keywords for each emotion, using WordNet to further expand these sets with synonyms. For each text sequence that it receives as input, the method calculates an individual score for each emotion, by adding one point every time it encounters a 'for' keyword relative to that emotion and subtracting one point every time it encounters an 'against' keyword. Once it has obtained a score for all four emotions (happiness/content, sadness, anger and anxiety/fear, computed in this order), it returns as output the highest-scoring emotion. In case of a tie, the emotion corresponding to the least index in the ordered list is selected among those whose score is highest.

Emotion classification experiments

We fine-tune two different language models (T5 and RoBERTa) for the task of emotion classification. For each model, we set up two distinct pipelines:

1. We apply one single fine-tuning to the model, using the subset of emotional expressions in the `EmpatheticPersonas` dataset.
2. We first fine-tune the model on the Emotion dataset (27). This dataset contains labelled expressions of emotion collected from Twitter, and it is fairly similar to our own corpus except for the fact that it includes two additional emotions (love and surprise), which we remove. We then apply a second fine-tuning to the model using our data.

Both datasets are split into train, validation and test sets in 80-10-10 proportions. We tune each model's hyperparameters using the train and validation sets (refer to Appendix D for details) and select the best performing combination. Each selected model is ultimately evaluated on the held-out test set obtained from our corpus.

Model evaluation

model	**accuracy**	**macro-f1**
baseline (rule-based)	63.03%	62.48%
T5 (single fine-tuning)	93.28%	93.51%
T5 (double fine-tuning)	93.28%	93.41%
RoBERTa (single fine-tuning)	89.92%	90.26%
RoBERTa (double fine-tuning)	**94.96%**	**95.10%**

Table 1: Results of the emotion recognition task. We observe significant improvement over the baseline, with the double fine-tuned RoBERTa model giving the best metrics.

All language models perform significantly better than the baseline, obtaining higher accuracy and macro-averaged f1 scores, regardless of whether they have received a double or single fine-tuning. The best-performing model, and thus the one chosen for our implementation, is the RoBERTa classifier fine-tuned on the Emotion dataset and then on our corpus, which achieves 94.96% accuracy and a macro-averaged f1 score of 95.10%. Table 1 compares the results achieved by each model.

5.3 Corpus augmentation

We notice that the utterance rewritings in the `EmpatheticPersonas` dataset consist of either one, two or three distinct sentences, separated by a full stop, a question mark or an exclamation point. Figure 2 illustrates an example of an utterance in our dataset that is composed of three sentences.

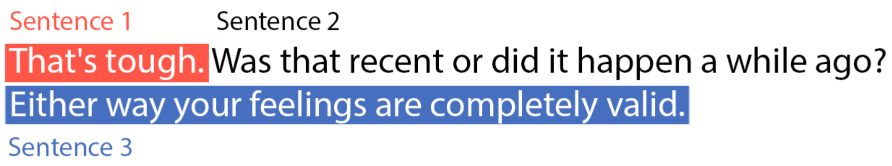

Figure 2: A rewritten utterance in the `EmpatheticPersonas` dataset, composed of three sentences. Sentence 2 conveys the main question, while Sentences 1 and 3 reinforce the empathy of the message by expressing sympathy and compassion.

We decide to extract all the individual sentences present in our dataset by splitting each utterance at major punctuation marks (full stops, questions marks and exclamation points). These sentences can then be recombined together in different ways to form new utterances. We believe this approach has several advantages: (a) it allows the augmentation of our text data, otherwise bound to the limited size of the dataset; (b) it ensures that the newly-generated utterances remain safe and reliable, since each sentence is self-contained in its meaning, has been reviewed at the dataset collection stage and is known not to be insensitive or harmful; (c) it has the potential to increase the level of empathy of those rewritten utterances which may not be highly empathetic in their original form. As shown in Figure 3, further analysis on our data shows that utterances composed of two or more sentences are perceived on average as more empathetic by human annotators compared to single-sentence ones. This may be due to the fact that, when an utterance is composed of several sentences, one of them conveys the main message while the others are often expression of politeness, sympathy or compassion.

When extracting sentences, we wish to save a record of their relative position within the original utterance. In this way, we can maintain this position when combining them together to form new utterances, thus increasing the likelihood that the result remains meaningful. To do so, we define three lists – $first\_pos\_list$, $second\_pos\_list$ and $third\_pos\_list$ – corresponding to the three possible positions within an utterance (since the utterances in our corpus contain at most three sentences), and assign each extracted sentence to one of these lists. Of course, this assignment is straight forward when an utterance is composed of three sentences, whereas for shorter utterances we need a strategy to achieve the most sensible assignment. The strategy that we use to populate with sentences the three position lists is illustrated in Algorithm 1.

Algorithm 1: Populate position lists

```
initialise empty lists first_pos_list, second_pos_list, third_pos_list
for each utterance u do
    initialise empty list temp_list
    split u into sentences at major punctuation marks
    append sentences to temp_list
if length of temp_list == 3 then
    // straight-forward case where 3 sentences were found in the
       utterance
    append first element of temp_list to first_pos_list
    append second element of temp_list to second_pos_list
    append third element of temp_list to third_pos_list
else
    if length of temp_list == 2 then
        // if 2 sentences were found, check whether the first is a
           question to determine where to assign them
        if first element of temp_list contains '?' then
            append first element of temp_list to second_pos_list
            append second element of temp_list to third_pos_list
        else
            append first element of temp_list to first_pos_list
            append second element of temp_list to second_pos_list
    else
        // if only once sentence was found, we assign it to the second
           position list
        append first element of temp_list to second_pos_list
```

After populating the position lists, we eliminate any duplicate sentences that they may contain. We also add an empty string to $first\_pos\_list$ and $third\_pos\_list$ (but not to $second\_pos\_list$, to which we assigned the sentences most likely to convey the main message of an utterance). It is therefore possible, by picking elements from each of these lists, to form new utterances that contain one, two or three sentences (but never empty utterances).

Since each new utterance is formed by successively choosing one item from $first\_pos\_list$, one item from $second\_pos\_list$, and one item from $third\_pos\_list$ until all possibilities have been exhausted, the resulting corpus will contain $|first\_pos\_list| \times |second\_pos\_list| \times |third\_pos\_list|$ utterances (where the notation $|list|$ indicates the length of $list$). It is worth noting that we repeat this process for each column in the `EmpatheticPersonas` dataset (i.e. we only combine together sentences originating from rewritings of the same base utterance).

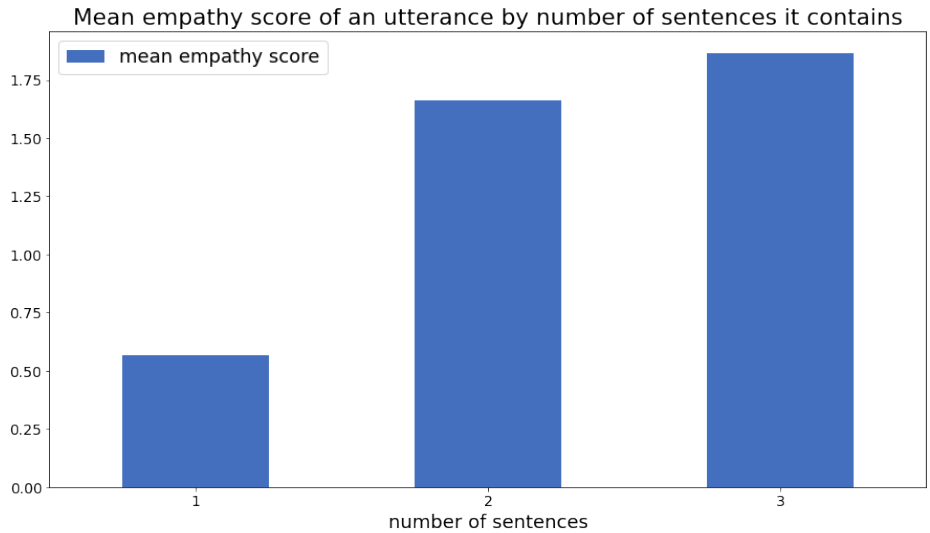

Figure 3: Bar chart showing the mean empathy score of utterances in the `EmpatheticPersonas` dataset according to the number of sentences that they contain, computed using the portion of the dataset that was annotated for empathy by human volunteers (for details of the annotation process see Subsection 3.4).

Following the process of sentence extraction and recombination, we obtain a corpus of utterances that is significantly larger than the original, as illustrated in Table 2.

Dataset split and associated persona	Total number of utterances before augmentation	Total number of utterances after augmentation
Males 40-69 (Robert)	480	3,980
Females 40-69 (Gabrielle)	495	4,123
Males 18-39 (Arman)	614	4,747
Females 18-39 (Olivia)	554	5,172
Entire dataset (Kai)	2,143	94,993

Table 2: Comparison of the total number of utterances in each dataset split before and after the augmentation process.

Visual inspection of the augmented corpora reveals that the quality of the newly-generated utterances is, on average, satisfactory. However, not all utterances are equally suitable to be used by the chatbot. Some of them may sound less fluent than others, due to repetitions or minor conflicts arising from putting together parts of different rewritings, and some may still lack enough empathy. Moreover, many utterances have sentences in common, increasing the risk that the bot's conversation may sound repetitive. To overcome these issues, we devise an appropriate retrieval method that yields the best possible utterance at each stage of the conversation. Our method consists of a multi-objective optimisation function combining an empathy score, a fluency score and a novelty score, which are simultaneously maximised when selecting an utterance.

5.4 Empathy score

To compute the empathy score of an utterance, we train a range of classifiers to predict the level of empathy of a text sequence and compare their performance.

Baseline

We use logistic regression over tf-idf as our empathy classification baseline, since this method has been shown to give good results when applied to similar problems (69; 70; 71).

Empathy classification experiments

We fine-tune a T5 and a RoBERTa model for the task of empathy classification. Once again, we build two distinct pipelines for each model:

1. We fine-tune the model on the portion of the `EmpatheticPersonas` dataset that has been annotated for empathy (see Subsection 3.4).

2. We apply to the model two consecutive fine-tunings. First, we train the model on the EPITOME dataset (44), which contains Reddit posts labelled with a discrete empathy score from 0 to 2, and then we further fine-tune its parameters on the empathy-annotated portion of our own dataset, whose labels are also discrete between 0 and 2.

We split both datasets into train, validation and test sets in 80-10-10 proportions, and select the best model in each category by tuning the hyperparameters using the train and validation sets (refer to Appendix D for details of the tuning process). We experiment with feeding into each model both unpreprocessed and preprocessed input sequences (lower-casing and removing punctuation in the latter). We evaluate each selected model on the `EmpatheticPersonas` test set.

Model evaluation

It is worth noting that empathy classification is a significantly harder task than emotion recognition, as evidenced by the fact that even the human annotators of our corpus are rarely in agreement when scoring the level of empathy of an utterance. The test results of all the models are compared in Table 3.

model	accuracy	macro-f1
baseline (tf-idf + logistic regression)	65.77%	65.78%
T5 (single fine-tuning on unpreprocessed data)	**80.18%**	**80.66%**
T5 (single fine-tuning on preprocessed data)	72.97%	73.37%
T5 (double fine-tuning on unpreprocessed data)	75.68%	75.35%
T5 (double fine-tuning on preprocessed data)	73.87%	74.22%
RoBERTa (single fine-tuning on unpreprocessed data)	73.87%	74.72%
RoBERTa (single fine-tuning on preprocessed data)	73.87%	74.95%
RoBERTa (double fine-tuning on unpreprocessed data)	74.77%	75.73%
RoBERTa (double fine-tuning on preprocessed data)	75.68%	76.11%

Table 3: Results for the empathy classification task.

All the pre-trained language model show an improvement over the baseline. The best performing model is the T5 classifier fine-tuned on the `EmpatheticPersonas` train set alone, which achieves an accuracy of 80.18% and a macro-averaged f1 score of 80.66%. Therefore, we select this model to be our empathy scoring function E. The values returned by this function should be normalised between 0 and 1, and thus we divide each output by 2 (which is the maximum empathy score possible).

5.5 Fluency score

In order to evaluate the fluency of an utterance, we compute the inverse of the perplexity (PPL) score returned by a GPT-2 language model. Since combining together portions of different utterances may create unwanted repetitions, we subtract from this value a penalty of 10^{-2} for each repeated (lemmatised) word, excluding stop words. Therefore, the fluency F of an utterance u is given by

$$F(u) = \frac{1}{\mathrm{PPL}(u)} - \mathrm{RP}(u),$$

where $\mathrm{RP}(u)$ is the total penalty for all the repeated words within that utterance. To normalise the fluency function so that it returns values through the whole range between 0 and 1, we divide $F(u)$ by the maximum possible fluency score calculated on our augmented corpus (0.16). If the output is negative, which may happen when the total penalty is greater than the inverse of the perplexity, we return zero.

5.6 Novelty score

Our chatbot is capable of saving and retrieving up to 50 of its previous utterances, and it compares each new utterance to all the utterances in this set to evaluate how novel it is. To this end, we implement a function that calculates the weighted overlap distance over all possible n-grams between two text sequences.

The overlap distance is obtained by subtracting the overlap coefficient, a measure of how similar two finite sets are (72), from the number 1. Given two sets A and B, their overlap distance (OD) is given by

$$\mathrm{OD}(A, B) = 1 - \frac{|A \cap B|}{\min(|A|, |B|)},$$

where the notation $|X|$ indicates the size of set X. This metric returns 0 when one of the two sets is a subset of the other, and 1 when the two sets have no elements in common.

We compute the overlap distance over all possible sets of n-grams contained in the two utterances that we are comparing, starting from unigrams up to N-grams where N is equal to the total number of tokens in the shorter of the two sentences, and sum the results. Of course, the greater the number n, the more we wish to weight the distance between n-grams, since utterances are more similar when they share longer sequences of words. For this reason, each overlap distance between n-grams is raised to the power n in our distance function. Finally, to ensure that the distance value over all n-grams remains between 0 and 1, we divide the resulting sum by N.

Given two utterances $u1$ and $u2$, their distance d is thus calculated as follows:

$$d(u1, u2) = \frac{\sum_{n=1}^{N} [\mathrm{OD}(n\text{-grams}(u1), n\text{-grams}(u2))]^n}{N},$$

where n-grams(u) represents the set of n-grams in the utterance u.

This metric is computed between a new utterance and each of the saved previous utterances, adding up the results to obtain the novelty (or diversity) score D of the new utterance. Once again, we divide D by the number of previous utterances to obtain a normalised value between 0 and 1.

5.7 Multi-objective optimisation function

Let $E_{norm}(u)$, $F_{norm}(u)$ and $D_{norm}(u)$ be the normalised functions measuring the empathy, fluency and diversity of an utterance u, each returning a value between 0 and 1. Then, the overall function R that we wish to maximise when retrieving a new utterance is given by

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

After extensive testing using different combinations of values, we fix the weights in the above function to $w_e = 1$, $w_f = 0.75$ and $w_d = 2$, as these appear to give the best results when retrieving utterances. It should be noted that calculating the output of $R(u)$ is computationally expensive. As a trade-off between computation speed and size of the utterance retrieval pool, we apply this function on a random subset of 25 utterances drawn from the augmented corpus.

5.8 Conversation flow

The conversation flow of the chatbot is largely based on the previous implementation (28), with a few structural changes.

After a user has logged into the platform, the chatbot asks for their first name. This is collected so that the conversational agent will be able to refer to the user by their name during the conversation. Following this step, the bot asks the user to choose a persona between Kai, Robert, Gabrielle, Arman and Olivia. The user's selection informs which portion of the (augmented) data is loaded into the back-end. This process is illustrated in Figure 4.

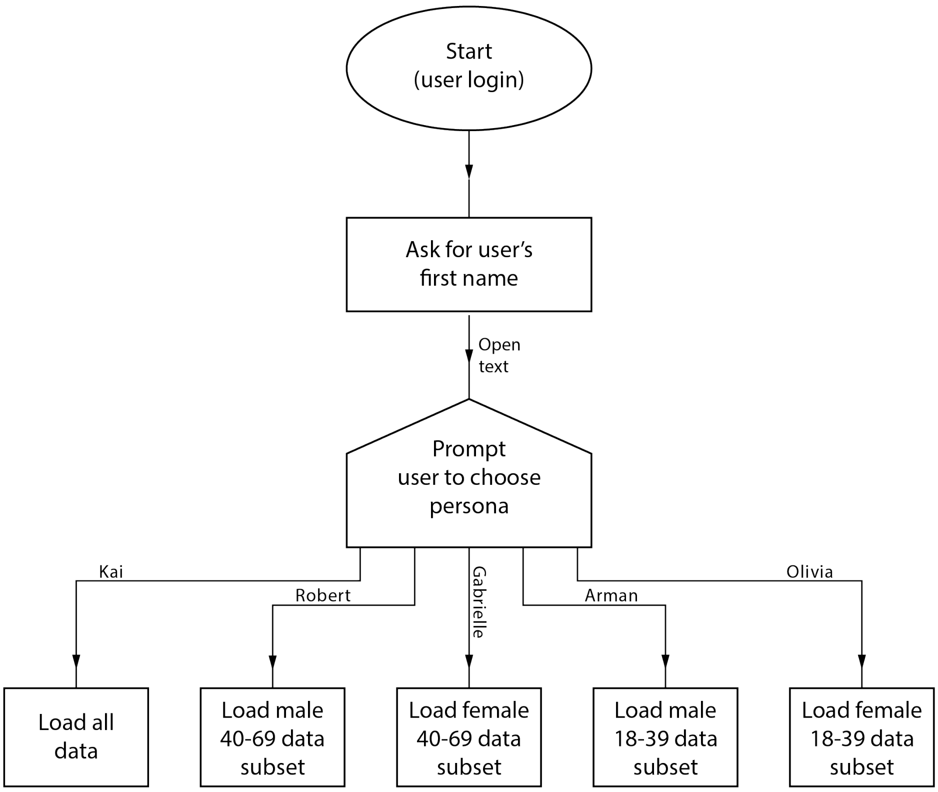

Figure 4: Having collected a user's first name and prompted them to choose a persona, the model loads the relevant portion of the augmented dataset.

The conversation flowchart has two main branches: one for positive emotions (happiness/content) and one for negative emotions (anger, sadness and anxiety/fear). Figures 5 and 6 illustrate both branches. Although the dialogue structure is the same for all negative emotions, the chatbot's utterances are different in each case and relevant to that specific emotional context (this is achieved by saving the user's emotion as a variable and selecting utterances from the dataset that are relevant to that emotion). Similarly, all five personas navigate the same flowchart when conversing with the user, but each of them has a specific set of utterances that they can choose from.

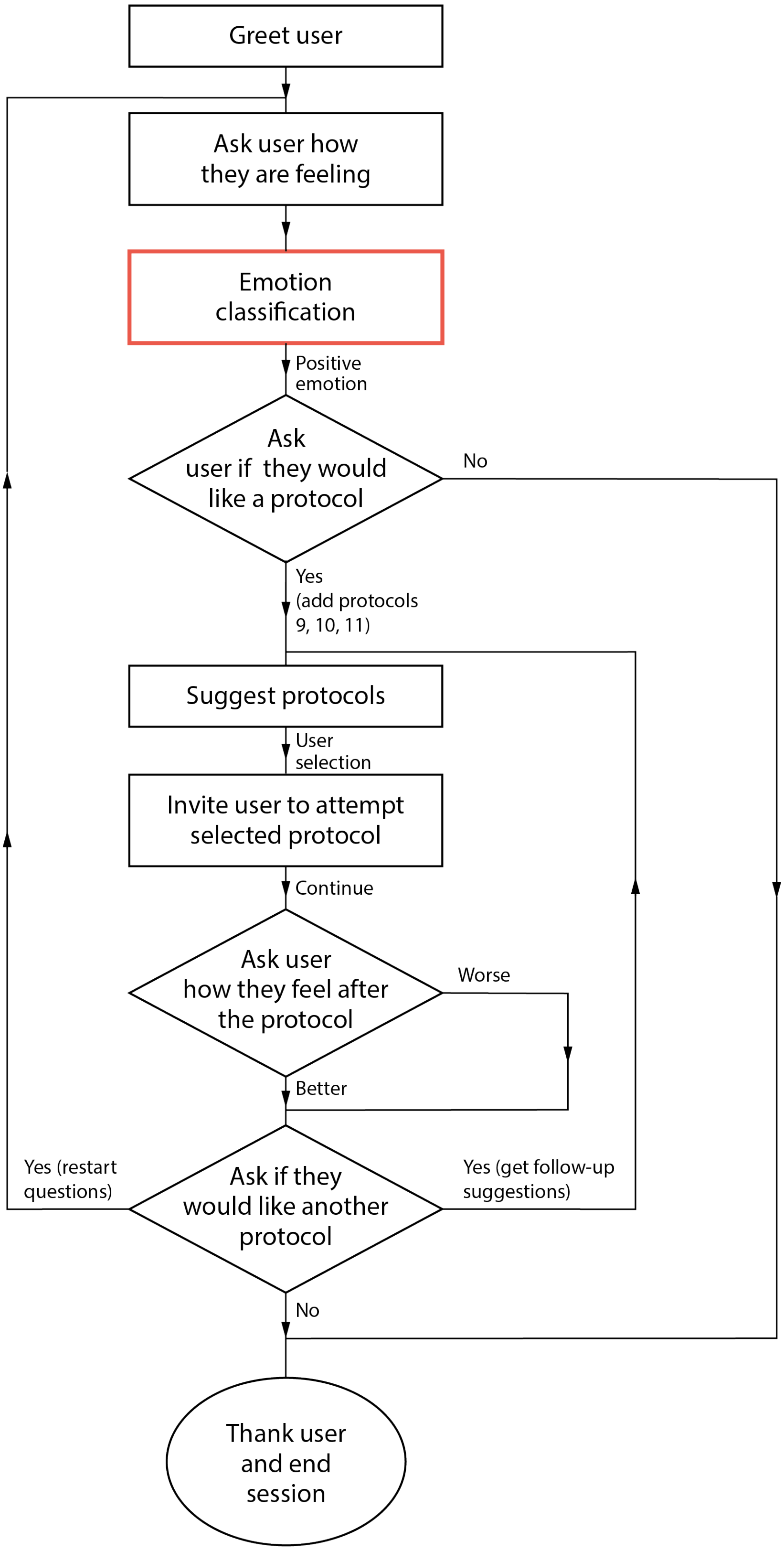

Figure 5: Conversation flow for positive emotions.

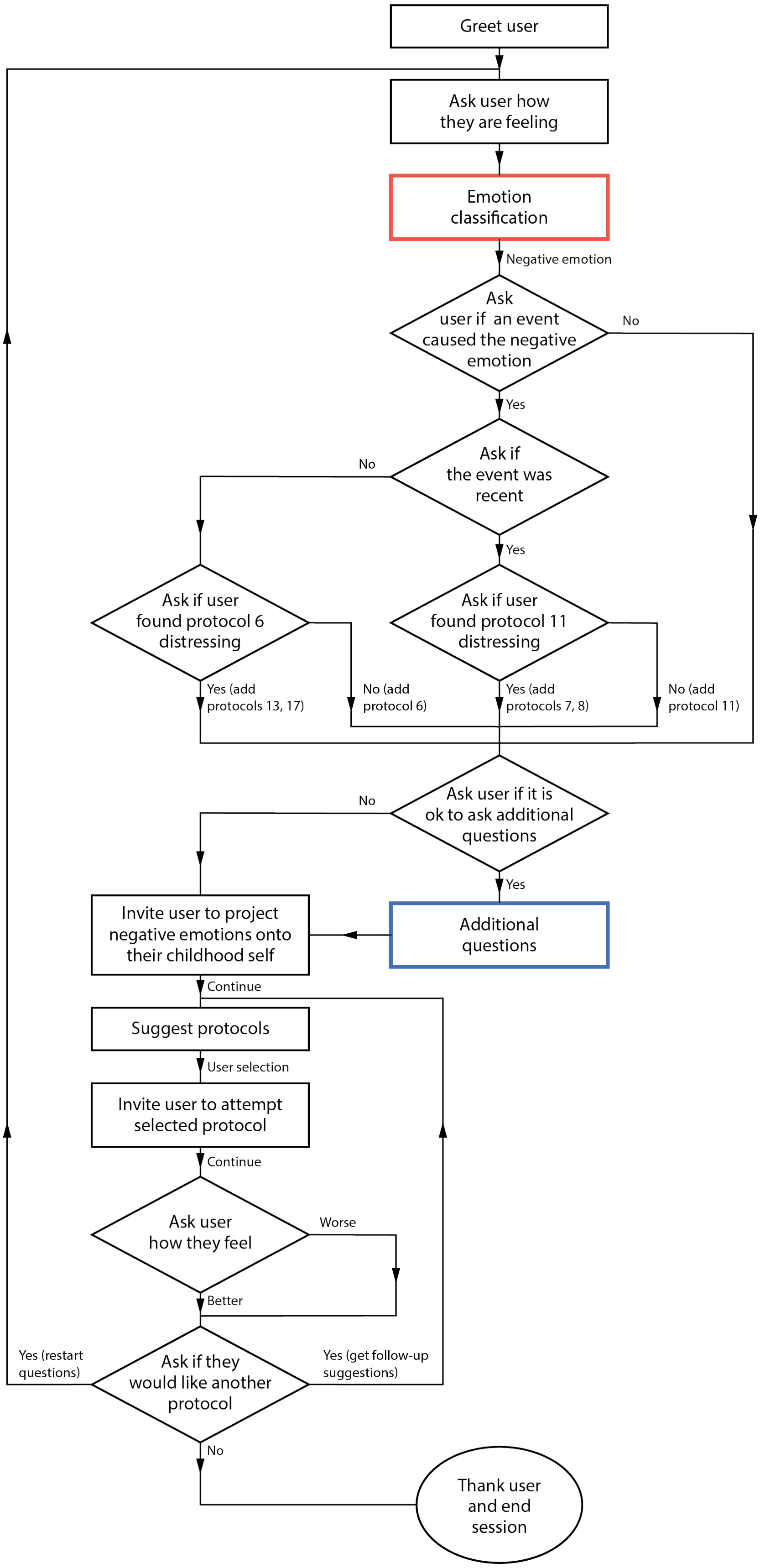

Figure 6: Conversation flow for negative emotions.

At the emotion classification stage (highlighted in red in Figures 5 and 6), the chatbot attempts to identify the user's emotion by passing their answer to the question 'How are you feeling?' into the classifier described in Subsection 5.2. This step is crucial to tailor the rest of the conversation to the relevant emotional context. To avoid any possibility of misclassification, the bot checks with the user whether the predicted emotion is correct, as illustrated in Figure 7. If the user confirms, it proceeds with the rest of the conversation, otherwise it asks them to select themselves from a list of emotions the one that most closely matches what they are feeling.

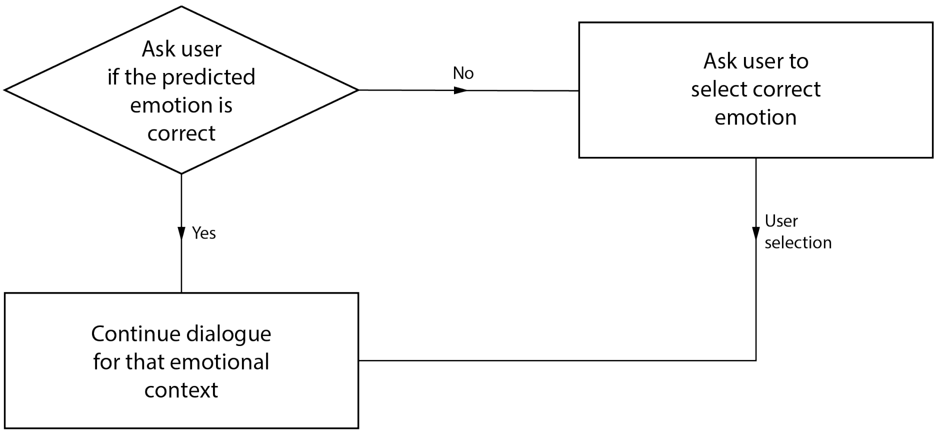

Figure 7: The chatbot verifies that they have correctly interpreted the user's emotional state before proceeding with the conversation.

To avoid as much as possible having conversations that are completely predetermined by the rigid structure of the flowchart, and thus potentially repetitive over time, a number of questions relating to the negative emotional context are randomly selected from a pool of options. Table 4 displays these questions and the protocol suggestions that answering each of them in a certain way would yield. Answering 'yes' to any of the questions prompts the bot to present the user with the protocol suggestions collected thus far, while answering 'no' leads to the random selection of another question from the pool. To prevent the chatbot from repeating the same type of question more than once within the same conversation, each of them is removed from the pool once it has been selected and asked. Figure 8 illustrates the general strategy followed by the bot to retrieve each of these questions, which is implemented at the 'additional questions' stage of the conversation flow, highlighted in blue in Figure 6. Of course, the questions are articulated differently depending on the selected persona as well as the specific emotional context of the user. In addition, like all the other utterances, each additional question has many variations, obtained applying the augmentation strategy described in Subsection 5.3 to the rewritings in the `EmpatheticPersonas` dataset.

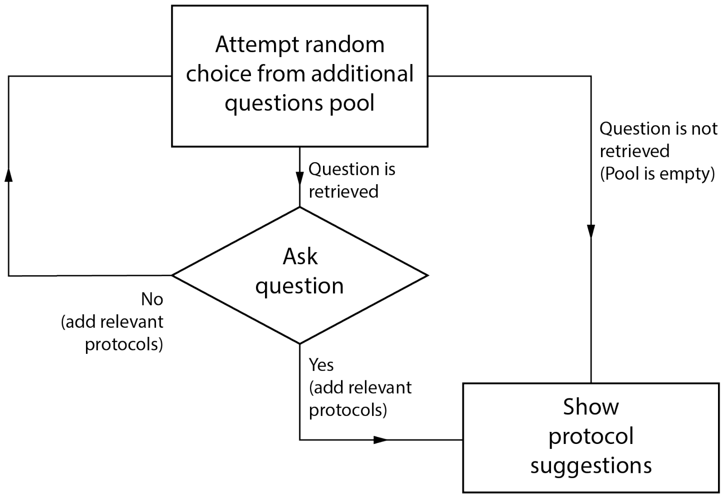

Figure 8: Random question retrieval strategy.

Base question	Behaviour after positive answer	Behaviour after negative answer
Have you strongly felt or expressed any of the following emotions towards someone: envy, jealousy, greed, hatred, mistrust, malevolence, or revengefulness?	Add protocols 13, 14, give suggestions	Add protocol 13, select next question
Do you believe that you should be the saviour of someone else?	Add protocols 8, 15, 16, 19, give suggestions	Add protocol 13, select next question
Do you see yourself as the victim, blaming someone else for how negative you feel?	Add protocols 8, 15, 16, 19, give suggestions	Add protocol 13, select next question
Do you feel that you are trying to control someone?	Add protocols 8, 15, 16, 19, give suggestions	Add protocol 13, select next question
Are you always blaming and accusing yourself for when something goes wrong?	Add protocols 8, 15, 16, 19, give suggestions	Add protocol 13, select next question
Is it possible that in previous conversations you may not have always considered other viewpoints presented?	Add protocols 13, 19, give suggestions	Add protocol 13, select next question
Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	Add protocols 13, 17, give suggestions	Add protocol 13, select next question

Table 4: Additional questions that can be asked by the chatbot in the context of negative emotions. Each question is retrieved from the list through a random choice, with the aim to make the chatbot's conversations more diverse.

5.9 User interface

Having created five distinct personas for our chatbot, we design an avatar for each of them. The personas created from subsets of responses in our dataset (Robert, Gabrielle, Arman, Olivia) are assigned quasi-photo-realistic avatars consistent with the sex and age range of the respondents that their dialogue is conditioned on. Kai – whose dialogue is informed by the entire `EmpatheticPersonas` dataset regardless of sex or age – is assigned a more abstract representation of a human face.

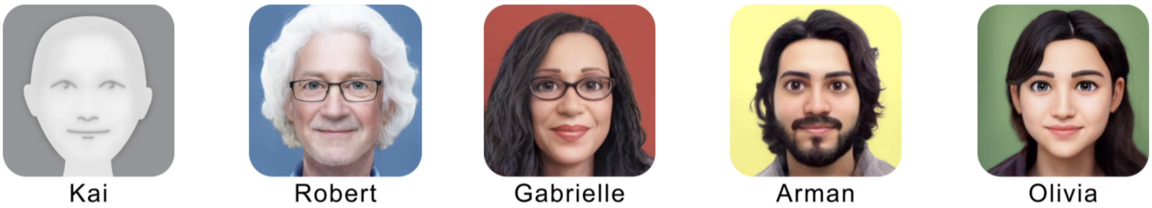

Figure 9: Avatars for the personas that SAT chatbot users can choose to interact with.

Similarly to the previous implementation of the SAT chatbot (28), the communication between the Python back-end and the JavaScript front-end is managed by the React-chatbot-kit library (73). The bot allows a mix of open text input and selection from predetermined options. The application is made available on the Web and is accessible via a computer or tablet. We embed into the interface a viewer to visualise the SAT protocols suggested by the bot. Figure 10 shows the appearance of the web application as it was set up for the evaluation trial.

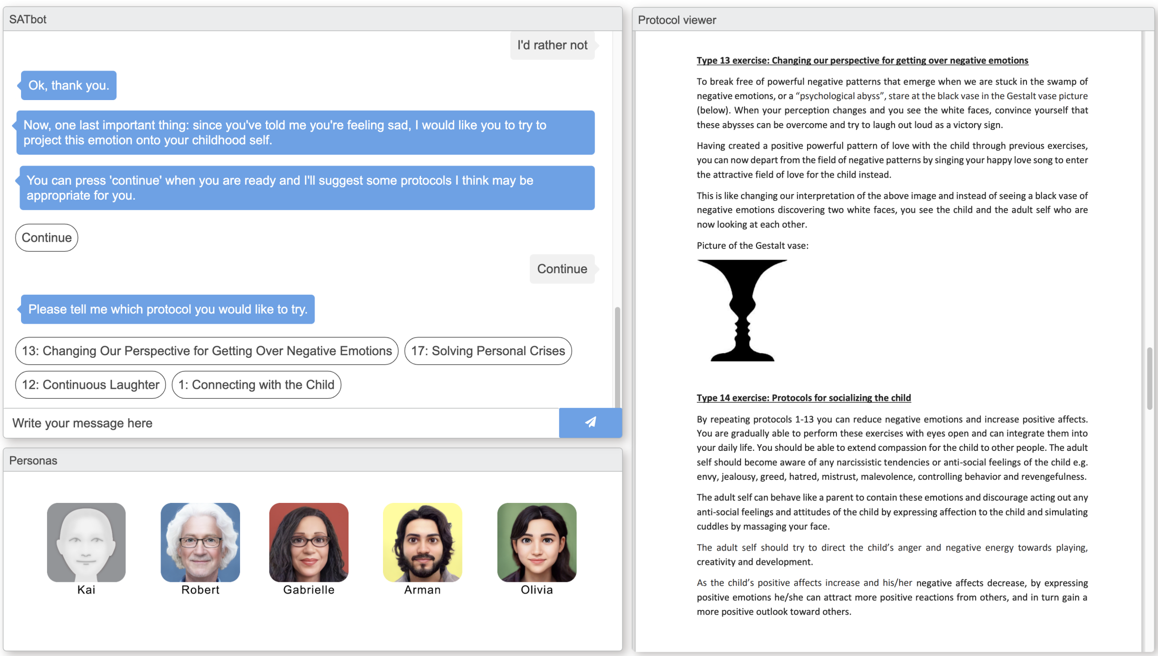

Figure 10: Appearance of the SAT chatbot web application.

6 Non-clinical trial

6.1 Study setup

The SAT chatbot is formally evaluated through a human trial. The pool of participants comprises 23 volunteers from the non-clinical population aged between 22 and 70, all of whom are already familiar with SAT psychotherapy and the SAT protocols. Of these 23 individuals, 16 are male and seven are female. Each participant has agreed to have four interactions with the chatbot over the course of five days; two with the persona named Kai and the rest with any two of the other personas.

The chatbot platform has been deployed as a web application for the trial, and thus all the interactions between the volunteers and the bot occur online. Participants are sent instructions, a link to access the platform and individual login credentials via e-mail. Each of them has the possibility to give feedback after completing all four interactions by filling out a questionnaire (see subsection 6.2).

Of the 23 individuals that volunteered to take part in the study, 16 have submitted a completed questionnaire at the time of writing. The evaluation in Section 7 is thus carried out on a sample size of 16.

We also invite to take part in the evaluation trial a medical professional specialised in mental health, who completes the same amount of interactions with the chatbot as the other participants. We collect feedback from the clinician through a separate questionnaire.

6.2 Evaluation questionnaire

By answering the evaluation questionnaire, study participants give feedback on their experience interacting with the SAT chatbot. In particular, the questionnaire aims to evaluate: (a) how well the bot interprets a user's emotion; (b) how empathetic it is; (c) how engaging it is; (d) how useful it is. When volunteers evaluate the bot for empathy and engagement, they are asked to score these attributes separately for the persona named Kai and the other personas. By collecting this information, we wish to evaluate whether a more human-like character – such as Robert, Gabrielle, Arman and Olivia – can increase the bot's perceived empathy and a user's level of engagement with it. On the other hand, we also assess whether having a much larger pool of utterances to choose from (and thus potentially more diversity in the responses), as is the case for Kai, provides a significant advantage.

The evaluation questionnaires distributed to the volunteers and the clinician are identical and, in the case of the volunteers, completely anonymous. Both are delivered online and contain multiple-choice questions requiring to indicate a level of agreement/disagreement with a series of statements, as well as open questions for providing additional feedback, comments and suggestions.

7 Evaluation

7.1 Principal findings

In this subsection, we analyse the evaluation questionnaire responses collected during and after the intervention. We compare the results with those obtained in a previous evaluation trial (28) of the earlier implementation of the SAT chatbot (where available), which we define as our baseline, and draw conclusions. Appendix E includes an exhaustive overview of all the feedback received.

Evaluation by study participants

Trial participants are asked to evaluate the chatbot's ability to convey empathy by expressing how much they agree/disagree with the statement 'The chatbot displayed empathy in its responses throughout the conversation', first in the context of their interactions with Kai, and then in relation to any of the other personas. When interacting with Kai, 75% of participants agree that the chatbot displays empathy, while the remaining quarter select 'Strongly agree' and 'Neither agree nor disagree' in equal proportions. When the interactions are with any of the other personas, 56% agree that the bot is empathetic, 19% strongly agree and a quarter neither agree nor disagree. In both cases, no volunteers disagree that the bot conveys empathy and we observe a significant improvement over the baseline: only 29% of those participating in the previous trial agreed that the earlier version of the chatbot was empathetic, with 42% disagreeing with this statement. Figure 11 shows a comparison between the empathy evaluation of the previous implementation and ours.

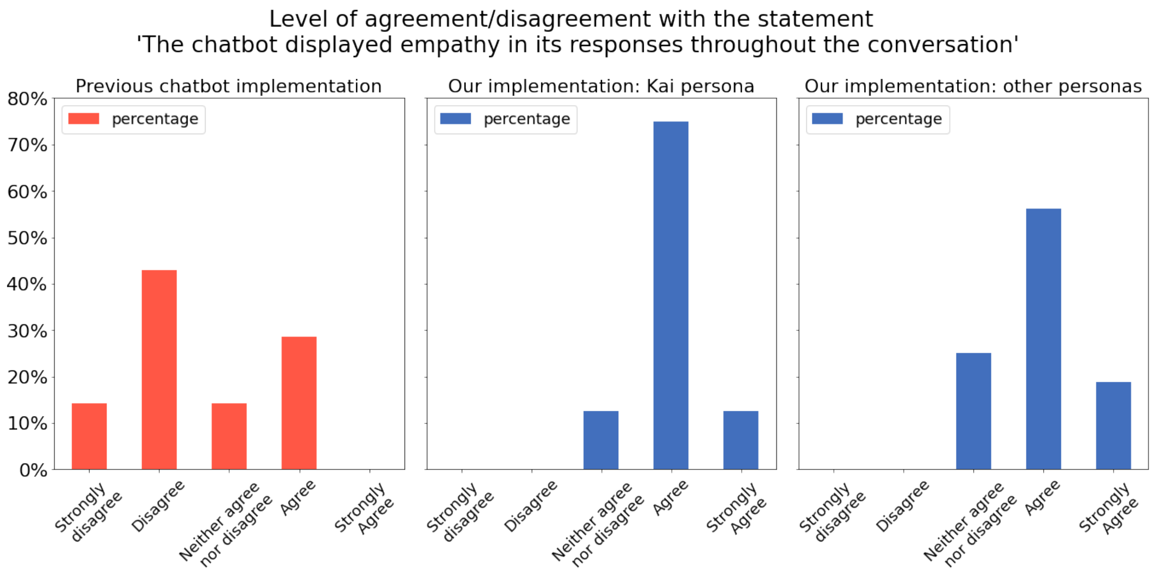

Figure 11: Empathy evaluation of the previous implementation of the chatbot (our baseline) and the current one. Our results show significant improvement over the baseline in the perceived level of empathy for both Kai and the other personas.

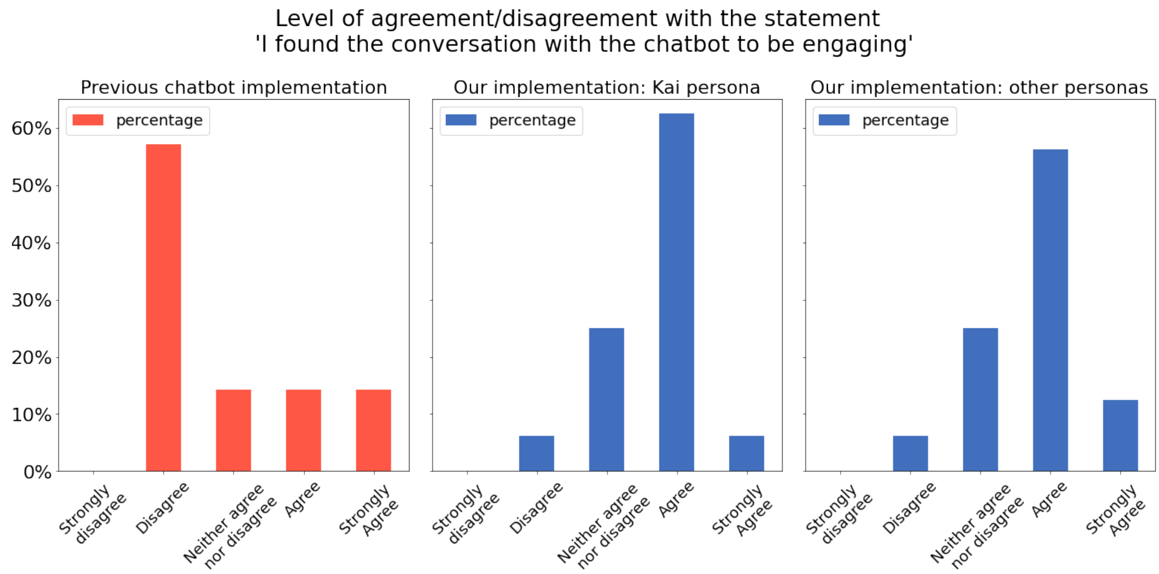

Figure 12: Engagement level of users who interacted with the previous version of the chatbot and ours. The comparison shows that the level of user engagement improves in our implementation, whether the interactions are with Kai or the other personas.

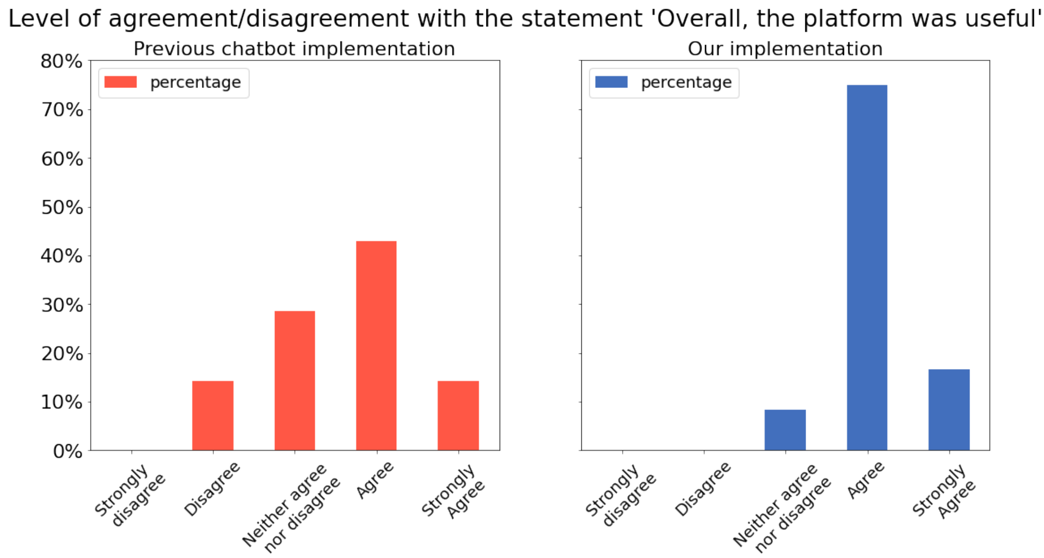

Figure 13: Evaluation of usefulness of the previous and current implementation of the SAT chatbot, showing that the current version is more consistently regarded as useful.

When evaluating their level of engagement with the platform, 6% of participants disagree with the statement 'I found the conversation with the chatbot to be engaging', and a quarter neither agree nor disagree. This is true for interactions with Kai as well as with the other personas. In the case of Kai, 63% agree that its conversations are engaging, and a further 6% strongly agree. On the other hand, 56% agree and 13% strongly agree that the other personas converse in an engaging manner. In comparison, when evaluating the previous implementation, over 57% disagreed that the dialogue was engaging, and the responses 'Strongly agree', 'Agree' and 'Neither agree nor disagree' were each selected by approximately 14% of participants. A comparison between the level of user engagement with the earlier chatbot version and ours is shown in Figure 12.

Overall usefulness of the application is evaluated by agreeing/disagreeing with the statement 'Overall, the platform was useful'. Of our sample, 77% agree and a further 15% strongly agree with the above statement, with 8% selecting the response 'Neither agree nor disagree'. No volunteers are in disagreement with this statement. As shown in Figure 13, this result is an improvement over the earlier version of the bot, where 43% agreed it was useful, 29% neither agreed nor disagreed, 14% strongly agreed and an equal proportion disagreed.

Regarding the ability of the chatbot to interpret a user's emotion, we find that 63% of the participants either agree or strongly agree with the statement 'The chatbot was good at guessing my emotion'. A quarter of them neither agree nor disagree, while the remainder disagree. Since no analogous data was collected during the previous trial, we cannot compare these results with the baseline. Instead, we refer the reader back to Subsection 5.2, where we compare our emotion classifier to the rule-based one implemented in the earlier chatbot version, and evaluate their results on a test set obtained from the `EmpatheticPersonas` dataset. Figure 14 illustrates the participants' evaluation of our bot's emotion recognition ability.

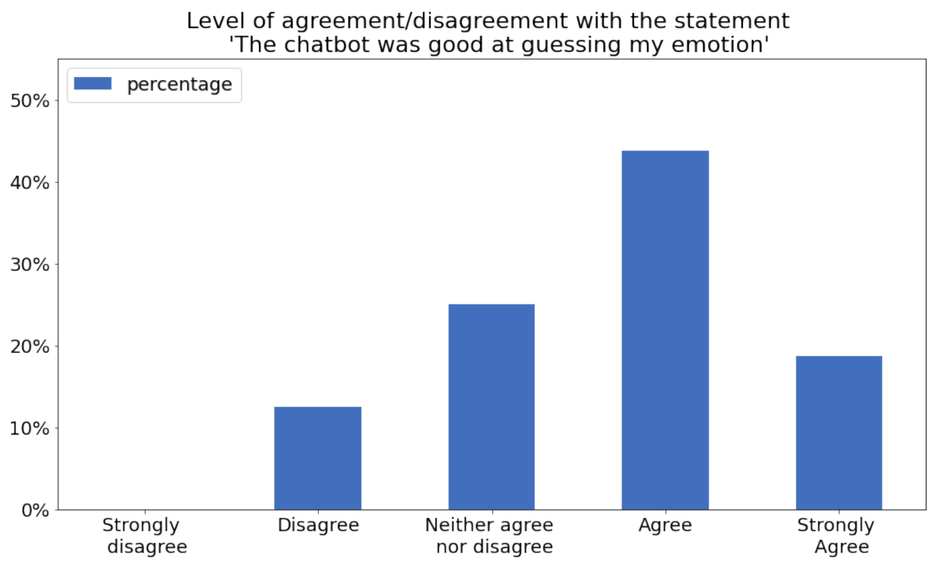

Figure 14: Evaluation of the current SAT chatbot's ability to recognise users' emotions.

Lastly, we investigate the volunteers' preferences when choosing personas, by asking them to state in the questionnaire who they have interacted with. Without considering the interactions with Kai (which were mandatory for all volunteers), we find that about a quarter of the remaining interactions were with Olivia, approximately 15% were with Gabrielle, and the remainder was equally split between Robert and Arman, who were chosen most often.

Overall, the chatbot is consistently rated higher than the baseline for empathy, usefulness and engagement. Its ability to recognise emotions earns a predominantly positive assessment. We should also note, however, the more critical feedback received through the questionnaires, in the form of comments and suggestions. Some participants observe that the range of feelings that the bot is able to recognise is too limited, and this affects its ability to successfully guess a user's emotional state. Further feedback around this topic highlights the fact that not only the range of emotions is narrow, but those emotions may also be quite extreme compared to what members of the non-clinical population might feel throughout the day. For example, feeling 'slightly worried' would be cast by the current version of the bot as being 'anxious/scared', whereas the two emotional states are arguably rather different.

While most volunteers welcomed the addition of personas, some of the comments indicate that their dialogue was at times perceived as too generic, and the questions were not in-depth enough. Relating to this, one participant suggests that the conversation would feel more natural if the bot only accepted open text responses. Two volunteers also note that having only two choices ('I feel better' and 'I feel worse') for giving feedback after completing a protocol is too limiting, and more nuanced options would be required. Finally, a number of suggestions revolve around changes in the user interface, particularly the necessity to have the bot available on mobile devices.

Clinician evaluation

Our framework is reviewed by one medical professional specialised in mental health, who has had four interactions with the chatbot (two with Kai, one with Gabrielle and one with Arman). The clinician's assessment of the current SAT chatbot and their evaluation of the previous version are shown in Table 5.

Statement	**Response (previous bot version)**	**Response (current bot version)**
The chatbot was good at guessing my emotion	N/A	Agree
The chatbot displayed empathy in their responses throughout the conversation	Disagree	Agree
I found the conversation with the chatbot to be engaging	Agree	Agree
Overall, the platform was useful.	Agree	Agree

Table 5: Clinician's evaluation of the previous and current version of the SAT chatbot.

Our implementation of the chatbot is viewed as significantly more empathetic by the clinician compared to the previous one, having turned their response to the statement 'The chatbot displayed empathy in their responses throughout the conversation' from 'Disagree' to 'Agree'. It should be noted that the clinician agrees with the above statement as well as with the statement 'I found the conversation with the chatbot to be engaging' both in the context of their interactions with Kai and in relation to their conversations with Gabrielle and Arman.

While the clinician is in agreement that the chatbot is good at guessing emotions and that it is empathetic, engaging and useful, they do not strongly agree with any of the statements. In the comment section of the questionnaire they explain that they perceived all of the personas' dialogue to be relatively uniform, and that their level of empathy was also very similar. They also observe that the empathy displayed by the bot is rather stereotypical and at times comes across as mechanical. Finally, they comment positively on the ability of the chatbot to interpret emotions, but note that this ability is limited by the narrow range of emotional contexts available.

7.2 Limitations of the study

Of 23 non-clinician volunteers that signed up to participate in the study, only 16 completed the evaluation questionnaire at the time of writing, resulting in a further reduction of an already modest sample. Moreover, since the questionnaires are completely anonymous, we have no way of knowing how the sex and age distribution of our actual sample (i.e. the sample of those who returned a completed questionnaire) compares to that of the entire pool of volunteers.

Ideally, the evaluation trial should be replicated with a larger sample of participants, and we should ensure that the distribution of this sample in terms of sex and age range (as well as any other characteristics that we may consider important for the purpose of the study) is not excessively skewed.

Of course, it should be noted that the small sample size is partly due to only enlisting participants who are already familiar with self-attachment technique. There is the risk that allowing individuals who have never practised or heard of SAT to take part in the study could bias the results, as their judgement of the chatbot may be coloured by their opinion of an unfamiliar therapeutic framework. Nevertheless, this is an issue that must be given careful consideration, as the small size and selected nature of the current sample prevent us from generalizing our findings.

In future trials, increasing the amount of required interactions and the length of the intervention may give participants a more informed opinion of the strengths and weaknesses of the chatbot, as some of these (e.g. its ability/inability to present the user with novel utterances over time) may only be evident over a period longer than five days. A longer intervention may also result in a higher proportion of participants giving feedback.

Fluency and novelty are two main objectives of our conversational framework, yet we evaluate them only indirectly by asking participants how engaging they found their conversations with the bot. We do this to be consistent with the questions posed during the previous trial, and thus have a dependable baseline for comparing our results. However, this leaves us with little insight into why a minority of the participants have found the bot not to be engaging. While the comment sections of the questionnaires clarify in part these results, it would be advisable to have the chatbot's dialogue explicitly evaluated for fluency and novelty in future studies.

8 Conclusions and future work

Our framework and study add to the existing body of knowledge in computational methods for mental health support. The human evaluation trial shows promising results with respect to the perceived empathy, usefulness, level of engagement and ability to interpret emotions of our application. Nevertheless, future work should address the issues and concerns that have been highlighted by trial participants in their feedback.

Firstly, despite achieving satisfactory metrics in our experiments, the emotion classifier integrated in the chatbot has room for improvement. Four emotional contexts are hardly enough to cover an acceptable range of human emotions. Collecting more data relative to different contexts as well as more nuanced feelings would be necessary to train a more competent model. It is also worth noting that the emotion expressions in the `EmpatheticPersonas` dataset have been provided by individuals instructed to answer the question 'How are you feeling?' *as if* they were experiencing a particular emotion. The fact that the model is not trained on genuine expressions of emotion – but rather on their imitation – is potentially a source of bias that may decrease its performance when applied in real-world situations.

The protocol recommendation mechanism could also be improved by expanding the pool of questions that the chatbot can ask before providing suggestions. Ideally, some if not all of these questions would imply an open text response from the user rather than requiring them to choose between predetermined answers, as this could increase the spontaneity of the communication. Of course, any model for the interpretation of open text input – whether keyword-based or relying on state-of-the-art deep learning techniques – carries the risk of misinterpretation. To reduce the possibility that the bot may misconstrue the user's input and carry out irrelevant conversations, appropriate measures should be put in place. One possible solution would be to include explicit checkpoints where the conversational agent asks for confirmation that their interpretation is correct – similarly to what happens at the emotion classification stage in the current implementation. However, it should be noted that having several of these checkpoints throughout the conversation could potentially break up the natural flow of the dialogue, nullifying any advantage deriving from an open-text format. Appropriate trade-offs between spontaneous conversations and reliable ones should be investigated, also taking into account potential differences in delivery style for the different personas.

When comparing the results obtained by the various chatbot personas during the evaluation trial, we find that participants report higher levels of user engagement when interacting with the human-like characters (Robert, Gabrielle, Arman, Olivia) than they do when they interact with Kai. While the overall percentage of approvals is the same for both types of persona, this proportion is significantly more skewed toward the top of the range (i.e. the 'Strongly agree' response) when users rate the former group. On the other hand, results are less conclusive when the chatbot is rated for empathy. In this context, the human-like personas still receive more top-of-the-range responses, however, when considering both 'Agree' and 'Strongly agree' answers, Kai is scored positively by a greater percentage of participants. To clarify these results and plan for future implementations, it may be necessary to repeat the study with a larger sample of volunteers. In order to design a more effective trial, the sex and age distribution of the participants should be considered carefully. For example, we have noted in this study that users favoured male characters over female ones (60% of all interactions were with Robert and Arman). This may or may not be due to the fact that males were over represented in our sample, and running the evaluation with a more evenly distributed sample could help validate or disprove this hypothesis.

One important thing to note is the significant computational cost of the multi-objective optimisation function that we use to retrieve utterances. Passing a single text sequence through a transformer model, which we use to score both empathy and fluency, has complexity $O(m^2)$, where m is the number of words within the sequence (60). Runtime is a crucial issue for an interactive chatbot; therefore, as explained in Subsection 5.7, we attempt to limit the computation cost by randomly selecting a sample of utterances from the corpus and running the multi-objective function on that sample only. It should be noted, however, that this may also limit the quality of the chosen utterance, since we have no way of predicting the overall fluency, empathy and novelty of the utterances in the sample. In future implementations, it could be worth precomputing the empathy and fluency scores of each utterance and appending these values to the augmented dataset. That way, only the novelty score, which depends on the bot's previous utterances, would need to be calculated in real time. The novelty function performs for each utterance $p \times N \times (N+1)/2$ comparisons, where p is the number of saved previous utterances and N is the length, in words, of the shorter of the two utterances being compared. This function is less computationally expensive than the other two when the number of previous utterances is small (however its complexity increases significantly as the conversation proceeds and more utterances are saved).

Finally, we believe that it is worth exploring further the trade-off between the safety and reliability offered by rule-based modelling and the variety of deep-learning informed dialogue. In this respect, we note that the `EmpatheticPersonas` dataset can be arranged into a parallel corpus of non-empathetic and empathetic utterances that have the same meaning, by using each base utterance as the non-empathetic version and every rewriting as the empathetic one. Having arranged the dataset in this way, it would be possible to attempt to train one or more sequence-to-sequence models for the task of monolingual translation of an utterance into its more empathetic interpretation. Should such a framework achieve satisfactory results, it could potentially be used to 'fix' the utterances generated by a language model by making them empathetic, and thus safer, before presenting them to a user. Of course, such an approach would be significantly more unpredictable than the current one and an appropriate assessment of the risks involved should be made before deploying it in a mental healthcare setting.

References

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. 2018;392(10159):1789–1858.

[2] Abbott A. COVID's mental-health toll: how scientists are tracking a surge in depression. *Nature*. 2021;590:194–195. Available from: `https://doi.org/10.1038/d41586-021-00175-z`.

[3] Nikcevic A, Spada M. The COVID-19 Anxiety Syndrome Scale: Development and psychometric properties. *Psychiatry Research*. 2020;292:113322. Available from: `https://doi.org/10.1016/j.psychres.2020.113322`.

[4] Lubker DKV. Socioeconomic Status and Domestic Violence. *International Journal of Global Health and Health Disparities*. 2004;3(1):85–91.

[5] Dobos N. Exploitation, Working Poverty, and the Expressive Power of Wages. *Journal of Applied Philosophy*. 2019;36(2):333–347. Available from: `https://doi.org/10.1111/japp.12314`.

[6] Lofstrom M, Raphael S. Crime, the Criminal Justice System, and Socioeconomic Inequality. *Journal of Economic Perspectives*. 2016;30(2):103–126. Available from: `https://doi.org/10.1257/jep.30.2.103`.

[7] Patel V, Lund C, Hatherill S, Plagerson S, Corrigall J, Funk M, et al. Mental disorders: equity and social determinants. In: Blas E, Sivasankara Kurup A, editors. *Equity, social determinants and public health programmes*. Geneva, Switzerland: World Health Organization Press; 2010. p. 115–134.

[8] Delgadillo J. Worlds apart: Social inequalities and psychological care. *Counselling and Psychotherapy Research*. 2018;18(2):111–113. Available from: `https://doi.org/10.1002/capr.12168`.

[9] Frankish H, Boyce N, Horton R. Mental health for all: a global goal. *Lancet*. 2018;392(10157):1493–1494. Available from: `https://doi.org/10.1016/S0140-6736(18)32271-2`.

[10] Fu Z, Burger H, Arjadi R, Bockting C. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. *Lancet Psychiatry*. 2020;7(10):851–864. Available from: `https://doi.org/10.1016/S2215-0366(20)30256-X`.

[11] Renn B, Hoeft T, Lee H, Bauer A, Arean P. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. *NPJ Digital Medicine*. 2019;2(1):6. Available from: `https://doi.org/10.1038/s41746-019-0077-1`.

[12] 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):300060520928686. Available from: https://doi.org/10.1177/0300060520928686.

[13] Gilburt H, Rose D, Slade M. The importance of relationships in mental health care: A qualitative study of service users' experiences of psychiatric hospital admission in the UK. *BMC Health Services Research*. 2008;8(1):92.

[14] Lambert M, Barley D. Research summary on the therapeutic relationship and psychotherapy outcome. *Psychotherapy: Theory, Research, Practice, Training*. 2001;38(4):357–361.

[15] World Bank Group. *World development report 2016 : digital dividends overview*. Washington D.C.; 2016. Annual report. Available from: http://documents.worldbank.org/curated/en/961621467994698644/World-development-report-2016-digital-dividends-overview [Accessed 11th May 2021].

[16] Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. *Behaviour Research and Therapy*. 2017;88:19–25. Available from: https://doi.org/10.1016/j.brat.2016.08.012.

[17] Blumenfield S, Levin-Scherz J. Digital Tools Are Revolutionizing Mental Health Care in the U.S. *Harvard Business Review*. 2020 Dec. Online article. Available from: https://hbr.org/2020/12/digital-tools-are-revolutionizing-mental-health-care-in-the-u-s [Accessed 18th August 2021].

[18] Gaffney H, Mansell W, Tai S. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Mental Health*. 2019;6(10):e14166. Available from: https://doi.org/10.2196/14166.

[19] Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Internal Medicine*. 2016;176(5):619–625.

[20] Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R, et al. Ethical Challenges in Data-Driven Dialogue Systems. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery; 2018. p. 123–129. Available from: https://doi.org/10.1145/3278721.3278777.

[21] Abd-alrazaq A, Alajlani M, Alalwan A, Bewick B, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*. 2019;132:103978. Available from: https://doi.org/10.1016/j.ijmedinf.2019.103978.

[22] Kapler J. Your chatbot lacks empathy, and that's a problem. *Venture Beat*. 2017 Jun. Online article. Available from: `https://venturebeat.com/2017/06/27/your-chatbot-lacks-empathy-and-thats-a-problem/` [Accessed 18th August 2021].

[23] Kislay K. Chatbots In Mental Health. Friendly But Not Too Friendly. *Analytics India Magazine*. 2020 Oct. Online Article. Available from: `https://analyticsindiamag.com/chatbots-in-mental-health-friendly-but-not-too-friendly/` [Accessed 18th August 2021].

[24] Raffel C, Shazeer N, Roberts A, , Lee K, Narang S, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. 2020;21:1–67.

[25] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*. 2019;1(8):9.

[26] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint. 2020. Available from: `https://arxiv.org/abs/1907.11692`.

[27] 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. 3687–3697. Available from: `https://doi.org/10.18653/v1/D18-1404`.

[28] Ghachem A. Evaluation of a Virtual Agent in Guiding Users from the Non-Clinical Population in Self-Attachment Intervention. Imperial College London MEng thesis. 2021.

[29] Kraus M, Seldschopf P, Minker W. Towards the Development of a Trustworthy Chatbot for Mental Health Applications. In: Lokoč J, Skopal T, Schoeffmann K, Mezaris V, Li X, Vrochidis S, et al., editors. *MultiMedia Modeling*. Cham, Switzerland: Springer International Publishing; 2021. p. 354–366.

[30] Denecke K, Vaaheesan S, Arulnathan A. A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. *IEEE Transactions on Emerging Topics in Computing*. 2020;PP(99):1. Available from: `https://doi.org/10.1109/TETC.2020.2974478`.

[31] Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*. 2017;10:39–46. Available from: `https://doi.org/10.1016/j.invent.2017.10.002`.

[32] Morbini F, Forbell E, DeVault D, Sagae K, Traum D, Rizzo A. A Mixed-Initiative Conversational Dialogue System for Healthcare. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Seoul, South Korea: Association for Computational Linguistics; 2012. p. 137–139.

[33] Bauer T, Devrim E, Glazunov M, Jaramillo WL, Mohan B, Spanakis G. #MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment. In: Cellier P, Driessens K, editors. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019. Communications in Computer and Information Science*. vol. 1167. Cham, Switzerland: Springer International Publishing; 2019. Available from: `https://doi.org/10.1007/978-3-030-43823-4_41`.

[34] 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 Mental Health*. 2017;4(2):e19. Available from: `https://doi.org/10.2196/mental.7785`.

[35] Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. Towards Understanding Emotional Intelligence for Behavior Change Chatbots. In: *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Los Alamitos, CA, USA: IEEE Computer Society; 2019. p. 8–14. Available from: `https://doi.org/10.1109/ACII.2019.8925433`.

[36] Ali MR, Razavi SZ, Langevin R, Al Mamun A, Kane B, Rawassizadeh R, et al. A Virtual Conversational Agent for Teens with Autism Spectrum Disorder. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 2020:1–8. Available from: `https://doi.org/10.1145/3383652.3423900`.

[37] Følstad A, Brandtzaeg P. Users' experiences with chatbots: findings from a questionnaire study. Quality and User Experience. 2020;5(3):1–14. Available from: `https://doi.org/10.1007/s41233-020-00033-2`.

[38] Edalat A. Introduction to self-attachment and its neural basis. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*; 2015. p. 1–8. Available from: `https://doi.org/10.1109/IJCNN.2015.7280780`.

[39] Ghaznavi I, Jehanzeb U, Edalat A, Gillies D. Usability evaluation of an immersive virtual reality platform for self-attachment psychotherapy. In: *CMH*; 2019. Available from: `https://dx.doi.org/10.475/123_4`.

[40] Elliott R, Bohart A, Watson J, Murphy D, Outcome E. Therapist Empathy and Client Outcome: An Updated Meta-Analysis. Psychotherapy. 2018;55(4):399–410. Available from: `https://doi.org/10.1037/pst0000175`.

[41] Kaluzeviciute G. The role of empathy in psychoanalytic psychotherapy: A historical exploration. *Cogent Psychology*. 2020;7(1):1748792. Available from: `https://doi.org/10.1080/23311908.2020.1748792`.

[42] Yalçin ÖN, DiPaola S. Modeling empathy: building a link between affective and cognitive processes. *Artificial Intelligence Review*. 2020;53(4):2983–3006. Available from: `https://doi.org/10.1007/s10462-019-09753-0`.

[43] Barrett-Lennard G. The Empathy Cycle: Refinement of a Nuclear Concept. *Journal of Counseling Psychology*. 1981;28:91–100. Available from: https://doi.org/10.1037/0022-0167.28.2.91.

[44] 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*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics; 2020. p. 5263–5276. Available from: https://doi.org/10.18653/v1/2020.emnlp-main.425.

[45] Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, et al. The Personalization of Conversational Agents in Health Care: Systematic Review. *Journal of Medical Internet Research*. 2019;21(11):e15360. Available from: https://doi.org/10.2196/15360.

[46] Landes S, Burton J, King K, Sullivan B. Women's Preference of Therapist Based on Sex of Therapist and Presenting Problem: An Analogue Study. *Counselling psychology quarterly*. 2013;26:330–342. Available from: https://doi.org/10.1080/09515070.2013.819795.

[47] Kessler EM, Rahn S, Klapproth F. Do young people prefer older psychotherapists? *European Journal of Ageing*. 2020;17:119–124. Available from: https://doi.org/10.1007/s10433-019-00519-9.

[48] Department for Digital, Culture, Media and Sport. *The Data Protection Act 2018 Keeling Schedule*. 2020. Act of Parliament. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969513/20201102_-_DPA_-__MASTER__Keeling_Schedule__with_changes_highlighted__V4.pdf [Accessed 25th May 2021].

[49] Department for Digital, Culture, Media and Sport. *General Data Protection Regulation Keeling Schedule*. 2020. Act of Parliament. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969514/20201102_-_GDPR_-__MASTER__Keeling_Schedule__with_changes_highlighted__V4.pdf [Accessed 25th May 2021].

[50] Amazon.com, Inc. *Amazon Mechanical Turk: access a global, on-demand, 24x7 workforce*. 2021. Web page. Available from: https://www.mturk.com/ [Accessed 20th August 2021].

[51] Prolific Academic Limited. *Quickly find research participants you can trust*. 2021. Web page. Available from: https://prolific.co/ [Accessed 20th August 2021].

[52] Luhn HP. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*. 1957;1(4):309–317. Available from: https://doi.org/10.1147/rd.14.0309.

[53] Jones K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 1972;28(1):11–21.

[54] Kuhn R, De Mori R. Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1990;12:570–583. Available from: https://doi.org/10.1109/34.56193.

[55] Rosenfeld R. A maximum entropy approach to adaptive statistical language modelling. *Computer Speech and Language*. 1996;10:187–228.

[56] Jurafsky D, Martin JH. N-gram Language Models. In: *Speech and Language Processing, 3rd edition*; 2020. p. 29–53.

[57] Jing K, Xu J. A Survey on Neural Network Language Models. arXiv preprint. 2019. Available from: https://arxiv.org/abs/1906.03591.

[58] Jordan MI. Chapter 25 - Serial Order: A Parallel Distributed Processing Approach. In: Donahoe JW, Packard Dorsel V, editors. *Neural-Network Models of Cognition*. vol. 121 of Advances in Psychology. Amsterdam, Netherlands: North-Holland Publishing; 1997. p. 471–495.

[59] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. vol. 1. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics; 2018. p. 2227–2237. Available from: https://doi.org/10.18653/v1/N18-1202.

[60] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–6010.

[61] Wang C, Li M, Smola AJ. Language Models with Transformers. arXiv preprint. 2019. Available from: https://arxiv.org/abs/1904.09408.

[62] Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*; 2018. p. 328–39.

[63] The Hugging Face Team. *Pretrained models*; 2020. Technical documentation. Available from: https://huggingface.co/transformers/pretrained_models.html [Accessed 18th August 2021].

[64] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT*; 2019. p. 4171–4186.

[65] Pradhan S, Moschitti A, Xue N, Uryupina O, Zhang Y. CoNLL-2012 shared task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In: *Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics; 2012. p. 1–40.

[66] Islam MA. *Comparing BERT and GPT-2 as Language Models to Score the Grammatical Correctness of a Sentence*. 2020 Oct. Weblog. Available from: `https://www.scribendi.ai/comparing-bert-and-gpt-2-as-language-models-to-score-the-grammatical-correctness-of-a-sentence/` [Accessed on 18th August 2021].

[67] Sun F, Belatreche A, Coleman S, Mcginnity TM, Li Y. Pre-processing Online Financial Text for Sentiment Classification: A Natural Language Processing Approach. In: *Proceedings of the 2014 IEEE Conference on Computational Intelligence for Financial Engineering Economics (CIFEr)*; 2014. p. 122–129. Available from: `https://doi.org/10.13140/2.1.3554.9443`.

[68] Camacho-Collados J, Pilehvar MT. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In: Linzen T, Chrupala G, Alishahi A, editors. *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics; 2018. p. 40–46.

[69] Vettigli G, Sorgente A. EmpNa at WASSA 2021: A Lightweight Model for the Prediction of Empathy, Distress and Emotions from Reactions to News Stories. In: *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics; 2021. p. 264–268.

[70] Kusumaningrum R, Indihatmoko T, Juwita S, Hanifah A, Khadijah K, Surarso B. Benchmarking of Multi-Class Algorithms for Classifying Documents Related to Stunting. *Applied Sciences*. 2020;10:8621. Available from: `https://doi.org/10.3390/app10238621`.

[71] Wang Y, Zhou Z, Jin S, Liu D, Lu M. Comparisons and Selections of Features and Classifiers for Short Text Classification. *IOP Conference Series: Materials Science and Engineering*. 2017;261:012018. Available from: `https://doi.org/10.1088/1757-899x/261/1/012018`.

[72] Vijaymeena MK, Kavitha K. A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications: An International Journal*. 2016 03;3(1):19–28. Available from: `https://doi.org/10.5121/mlaij.2016.3103`.

[73] Oseberg F. *React-chatbot-kit*; 2021. Technical documentation. Available from: `https://fredrikoseberg.github.io/react-chatbot-kit-docs/` [Accessed 22nd August 2021].

Appendices

A Overview of SAT protocols

We provide a brief overview of the 20 self-attachment technique protocols. Recommending the most appropriate protocols from this list is the main objective of the SAT chatbot.

1: Connecting with the Child

This protocol aims to encourage the patient to connect with their own childhood self. The patient tries to visualise the childhood self and imagine that they are cuddling them or playing with them. The connection can be aided by two childhood photographs of the patient, where these are available: one where they were happy and smiling and one where they were sad and frowning.

2: Laughing at our Two Childhood Pictures

The objective of this exercise is to teach the patient's childhood self to laugh. To achieve this, the patient laughs at their own childhood pictures and at the contrast between the opposite emotions that they express. It should be noted that this laughter is meant to make light of life and its events, never to ridicule the patient or their childhood self.

3: Falling in Love with the Child

Protocol 3 is aimed at establishing a deeper connection between the patient and their childhood self. The patient sings happy love songs out loud dedicating them to the childhood self. The patient is encouraged to increasingly use their whole body, dancing to the music as if they were dancing with their childhood self.

4: Vow to Adopt the Child as Your Own Child

In this exercise, the patient makes a life-long pledge to adopt their childhood self as their own child. They vow to support and care for the child by continuing to practise the self-attachment protocols. When making this pledge, the patient should be speaking out loud.

5: Maintaining a Loving Relationship with the Child

To nourish and maintain the relationship that they have established with the childhood self, patients select a short, loving phrase of their choice to repeat out loud. While focussing on their two childhood pictures, patients utter this phrase and recite love songs to the child. Again, patients are encouraged to use their whole body while doing so, moving and dancing to the melody of the song.

6: An exercise to Process the Painful Childhood Events

Protocol 6 aims to help patients process any difficult events that they may have endured during their childhood. In order to do so, a patient must try to visualise their childhood self experiencing the event in as much detail as possible, aided by their unhappy childhood photograph. They then must imagine their current, adult self comforting and embracing the distressed child, as a good parent would do. Any verbal interaction with the childhood self should preferably be spoken out loud, and the patient is encouraged to massage their own face while they picture themselves soothing and consoling the child.

7: Protocols for Creating Zest for Life

In this exercise, the patient looks as themselves in a mirror while dancing, reciting happy poems and singing love songs, imagining that what they see in the mirror is a reflection of their own childhood self. The exercise can also be repeated in different circumstances to aid its integration into daily life. For example, it can be carried out while doing housework.

8: Loosening Facial and Body Muscles

This protocol is designed to encourage patients to relax and loosen the muscles in their face and body, while imagining that they are interacting in a loving way with their childhood self.

9: Protocols for Attachment and Love of Nature

Patients are encouraged to maintain a beneficial relationship with nature and the outdoors, by visiting green areas located near where they are. Once they are in such a location, they should focus their attention on their surroundings, feeling admiration and wonder for the beauty of the natural world (for example by admiring a beautiful tree). This exercise should preferably be carried out at different locations over time.

10: Laughing at, and with One's Self

Protocol 10 celebrates an accomplishment – no matter how big or small – that the patient feels they have achieved. The patient is encouraged to smile about this accomplishment, and to gradually turn that smile into an authentic, wholehearted laughter. Over time, this laughter should last increasingly longer. By being taught to laugh at their own accomplishments, patients gradually learn to laugh at any life event. As already noted in the description of Protocol 2, laughter in self-attachment technique is compassionate, and it is never meant to ridicule the patient.

11: Processing Current Negative Emotions

In this protocol, patients project any negative feeling that they may be experiencing onto their childhood self, picturing the child in an unhappy state. They can then soothe the child's emotions by consoling and reassuring them, speaking calming words out loud. Once the patient has managed to overcome the unhappy feeling, they can imagine the child being content again, by looking at the happy childhood photograph or picturing it in their head.

12: Continuous Laughter

This exercise is aimed at managing and reducing stress. It requires the patient to find a quiet place and smile using all the muscles in their face (i.e. form a Duchenne smile). They can then vocalise out loud any or all of the following phrases, as if laughing: 'eh, eh, eh, eh'; 'ah, ah, ah, ah'; 'oh, oh, oh, oh'; 'uh, uh, uh, uh'; 'ye, ye, ye, ye'. The exercise also encourages patients to think of something funny to aid the laughter, be it an event that has happened or even just the silliness of the exercise itself.

13: Changing Our Perspective for Getting Over Negative Emotions

Protocol 13 is to be performed with the visual aid of a picture of the Gestalt vase. Looking at the picture, the patient tries to associate their switching perception of the image in it with changes in their attitude toward their problems. The black vase in the picture represents the patient's negative emotional patterns, whereas the two white faces – a symbol of the adult and childhood self now connected – are associated with victory in overcoming those patterns. Patients are encouraged to laugh victoriously when they manage to switch their perception of the image from the black vase to the white faces.

14: Protocols for Socializing the Child

In this exercise the patient, acting as their adult self, is invited to recognise any antisocial tendencies or behaviours of the inner child, and attempt to contain them and discourage them in an empathetic and compassionate manner. The negative energy fueling these tendencies is thus redirected toward more constructive and creative activities.

15: Recognising and Controlling Narcissism and the Internal Persecutor

Protocol 15 is a self-reflection exercise aimed at recognising and analysing one's own feelings and behaviours relating to persecution, victimhood and rescue, as well as their negative effects. In light of this analysis, the patient re-evaluates past experiences informed by these tendencies, and learns to identify them and avoid them.

16: Creating an Optimal Inner Model

This protocol focusses on identifying and recognising any emotions and tendencies that may have been developed during one's upbringing. Here, the adult self thinks about the behaviours that the childhood self formed during the early years, as a result of interactions with a parent or main carer. The adult self is then tasked with compassionately teaching the child to change and improve these behaviours.

17: Solving Personal Crises

Protocol 17 builds on the previous ones to help the patient resolve a crisis that they may be experiencing in their private life, such as a fall-out with friends or loved ones. First, the patient asks the childhood self whether this crisis originates from the negative behaviours addressed in Protocol 15, and whether it can be an opportunity for growth. Once again, the patient is invited to laugh at their problems, by practising Protocol 12 at the same time as they go through this exercise. As a next step, the adult self evaluates the situation from their own perspective, comparing it with past experiences and trying to learn from it.

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

In this exercise, the patient begins by reading out loud a quote from Nietzsche's *The Will to Power*, and is instructed to laugh while doing so. As they read through the quote, the patient is also invited to remember past negative experiences, and how these have made them into the person that they are today. After learning to laugh at distant trauma, patients can then gradually apply this exercise to recent and ongoing difficulties.

19: Changing Ideological Frameworks for Creativity

Protocol 19 encourages the patient to challenge their beliefs and convictions and learn to examine situations from multiple perspectives. The list of convictions that can be challenged include political leanings and personal ideas on cultural and social matters, such as issues pertaining race and sexuality. In each case, the patient debates these issues with themselves taking up the role of both proponent and opponent.

20: Affirmations

Patients are invited to draw a list of quotes, famous or otherwise, that they find powerful and which resonate with them. These quotes are to be read out loud whenever the patient needs to motivate themselves and find strength in the journey toward reaching their goals.

B Dataset summary

Tables 6, 7, 8 and 9 below illustrate the emotional contexts present in each 50-row subset of the `EmpatheticPersonas` dataset, as well as the number of emotion expressions and rewritings of each base utterance that they contain.

Row subset (indices)	**Context**	**Emotion expr. per context**	**Base utterances**	**Rewritings per base utterance**
0–49	Sadness	150	(Sadness) Was this caused by a specific event/s?	50
			(Sadness) Was this caused by a recent or distant event (or events)?	47
			(Sadness) Have you recently attempted protocol 6 and found this reignited unmanageable emotions as a result from past events?	48
			(Sadness) Have you recently attempted protocol 11 and found this reignited unmanageable emotions as a result from past events?	48
			(Sadness) Thank you. Now I will ask some questions to understand your situation.	50
			(Sadness) Have you strongly felt or expressed any of the following emotions towards someone:	50
	Anger	149	(Anger) Was this caused by a specific event/s?	44
			(Anger) Was this caused by a recent or distant event (or events)?	44
			(Anger) Have you recently attempted protocol 6 and found this reignited unmanageable emotions as a result from past events?	42
			(Anger) Have you recently attempted protocol 11 and found this reignited unmanageable emotions as a result from past events?	42
			(Anger) Thank you. Now I will ask some questions to understand your situation.	45
			(Anger) Have you strongly felt or expressed any of the following emotions towards someone:	42

Table 6: Number of emotion expressions and rewritings of each base utterance included in rows 0-49 of the dataset, for the contexts of sadness and anger.

Row subset (indices)	Context	Emotion expr. per context	Base utterances	Rewritings per base utterance
50–99	Sadness	149	(Sadness) Do you believe that you should be the saviour of someone else?	48
			(Sadness) Do you see yourself as the victim, blaming someone else for how negative you feel?	48
			(Sadness) Do you feel that you are trying to control someone?	47
			(Sadness) Are you always blaming and accusing yourself for when something goes wrong?	49
			(Sadness) In previous conversations, have you considered other viewpoints presented?	47
			(Sadness) Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	49
	Anger	148	(Anger) Do you believe that you should be the saviour of someone else?	44
			(Anger) Do you see yourself as the victim, blaming someone else for how negative you feel?	43
			(Anger) Do you feel that you are trying to control someone?	43
			(Anger) Are you always blaming and accusing yourself for when something goes wrong?	45
			(Anger) In previous conversations, have you considered other viewpoints presented?	44
			(Anger) Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	42

Table 7: Number of emotion expressions and rewritings of each base utterance included in rows 50-99 of the dataset, for the contexts of sadness and anger.

Row subset (indices)	**Context**	**Emotion expr. per context**	**Base utterances**	**Rewritings per base utterance**
100–149	Anxiety/ Fear	141	(Anxiety/ Fear) Was this caused by a specific event/s?	49
			(Anxiety/Fear) Was this caused by a recent or distant event (or events)?	50
			(Anxiety/Fear) Have you recently attempted protocol 6 and found this reignited unmanageable emotions as a result from past events?	49
			(Anxiety/Fear) Have you recently attempted protocol 11 and found this reignited unmanageable emotions as a result from past events?	49
			(Anxiety/Fear) Thank you. Now I will ask some questions to understand your situation.	50
			(Anxiety/Fear) Have you strongly felt or expressed any of the following emotions towards someone:	50
	Joy/ Content	150	N/A	
	All emotions	N/A	(All emotions) Here are my recommendations, please select the protocol that you would like to attempt	49
			(All emotions) Please try to go through this protocol now. When you finish, press ‘continue’.	49
			(All emotions) Do you feel better or worse after having taken this protocol?	49
			(All emotions) Would you like to attempt another protocol? (Patient feels better)	50
			(All emotions) Would you like to attempt another protocol? (Patient feels worse)	50

Table 8: Number of emotion expressions and rewritings of each base utterance in rows 100-149 of the dataset, for the contexts of anxiety/fear, joy/content and ‘all emotions’.

Row subset (indices)	Context	Emotion expr. per context	Base utterances	Rewritings per base utterance
149–199	Anxiety/Fear	144	(Anxiety/Fear) Do you believe that you should be the saviour of someone else?	49
			(Anxiety/Fear) Do you see yourself as the victim, blaming someone else for how negative you feel?	50
			(Anxiety/Fear) Do you feel that you are trying to control someone?	50
			(Anxiety/Fear) Are you always blaming and accusing yourself for when something goes wrong?	50
			(Anxiety/Fear) In previous conversations, have you considered other viewpoints presented?	50
			(Anxiety/Fear) Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	50
	Joy/Content	150	(Joy/Content) That's Good! Let me recommend a protocol you can attempt.	49
	All emotions	N/A	(All emotions) From what you have said I believe you are feeling {}. Is this correct?	50
			(All emotions) I am sorry. Please select from the emotions below the one that best reflects what you are feeling:	50
			(All emotions) Thank you for taking part. See you soon.	50

Table 9: Number of emotion expressions and rewritings of each base utterance in rows 150-199 of the dataset, for the contexts of anxiety/fear, joy/content and 'all emotions'.

C Sample conversations

Below are two sample conversations between the SAT chatbot and a human.

Conversation 1. Interlocutor: Kai, emotional context: happiness/content.

Kai:	Hello, this is Kai. How are you feeling today?
User:	Nothing is bothering me, I am quite relaxed
Kai:	Thank you for sharing your emotions. If I'm understanding correctly, you would identify as being happy/content, is that a fair assessment?
User:	Yes
Kai:	Oh, it's really great to hear it! Shall we pick a protocol together to keep this going? I hope it could inspire you and brighten your day even more.
User:	Okay
Kai:	I've got some ideas that might work. Which of the following would you like to explore?
User:	19: Changing Ideological Frameworks for Creativity
Kai:	You have selected Protocol 19. I advise you to attempt this protocol now, it may be difficult but just try your best and it will be fine. When you have finished, I'd be grateful if you could come back to me and press 'continue'.
User:	Continue
Kai:	Thank you for confirming that you have taken this protocol, I hope it was a positive experience for you. How would you say you feel now, better or worse?
User:	I feel better
Kai:	I'm so glad you found this helpful and you are feeling better having gone through it. Do you think you would also benefit from another protocol?
User:	No
Kai:	I appreciate that you've spent this time with me. I look forward to our next session.

Conversation 2. Interlocutor: Gabrielle, emotional context: anxiety/fear.

Gabrielle:	Hello Lisa, this is Gabrielle. How are you feeling today?
User:	Stressed out, I am not sure I will be able to make rent next month
Gabrielle:	Thank you, I just want to make sure that I have understood correctly. I would say that you're feeling anxious/scared right now, do you think that is a reasonable description of your current mood?
User:	Yes
Gabrielle:	That can be rough. So I can help you lessen your anxiety, are you able to think of a specific event that might have caused it?
User:	Yes, something happened
Gabrielle:	I'm sorry to hear that, I hope you don't mind elaborating a bit further. Was it a recent event or something that happened some time ago?
User:	It was recent
Gabrielle:	I can see why you may be feeling anxious. Have you had any experience with Protocol 11 and if so, did you find that it brought difficult emotions to the surface that were associated with recent events?
User:	No
Gabrielle:	Thank you very much for sharing this with me, I know that it is not an easy thing to do. I would like to ask some more questions now to understand the situation a little better.
User:	Okay
Gabrielle:	I hope you can trust me enough to let me know if you need help. Is there anything major happening in your life that might be causing you to be anxious?
User:	No
Gabrielle:	I really would like to understand more about what goes on in your life at the moment. Do you think you are responsible for someone else's happiness?
User:	No
Gabrielle:	Ok, I see. Could you perhaps be feeling like this because you see yourself as a victim of someone else's actions?
User:	No
Gabrielle:	Thank you for talking about this with me. Could it be that you are trying to control someone, without even realising it? Please rest assured that I would not judge you for it.
User:	No
Gabrielle:	Thank you, I'd like us to dig a little deeper if that's ok with you. Do you always take responsibility for everything negative that happens, even when it may not be your fault?

User:	Yes
Gabrielle:	Thank you, I will recommend some protocols for you in a moment. Before I do that, could you please try to project your anxious feeling onto your childhood self? Take your time to try this, and press ‘continue’ when you feel ready.
User:	Continue
Gabrielle:	Based on what you have shared with me, I feel that these recommendations would be appropriate for you. Can you let me know which protocol you’d like to try?
User:	11: Processing Current Negative Emotions
Gabrielle:	You have selected Protocol 11. That’s great, I’m glad you are willing to give this a try, hopefully you will find it very helpful. When you have completed it, press ‘continue’.
User:	Continue
Gabrielle:	I’d like to ask you about how that went. Has it made a difference, and if so do you feel better now, or worse?
User:	I feel worse
Gabrielle:	I’m sorry that hasn’t helped. If you would like to try another protocol to see if this suits you better, there are other options for you.
User:	No
Gabrielle:	I really appreciate the time that you have taken and all that you have shared with me. I hope we meet again soon.

D Hyperparameter tuning

T5

We search through the following hyperparameter space to fine-tune the T5 models for the tasks of emotion and empathy classification. In all cases we use the AdamW optimizer and monitor the validation loss during training to decide when to stop.

- **learning rate** = 1×10^{-5}, 5×10^{-5}, 1×10^{-4}, 5×10^{-5}, 1×10^{-3}
- **Adam epsilon** = 1×10^{-8}, 1×10^{-5}, 1×10^{-4}, 1×10^{-3}
- **gradient accumulation steps** = 1, 2, 4, 16
- **batch size** = 1, 4, 8, 16

RoBERTa

To fine tune the RoBERTa classifiers, we search through the hyperparameter space shown below. Again, we use the AdamW optimizer in all cases and stop training as soon as the validation loss starts rising.

- **learning rate** = 1.35×10^{-5}, 1.35×10^{-4}, 1.35×10^{-3}
- **Adam epsilon** = 1×10^{-8}
- **gradient accumulation steps** = 1, 2, 4
- **batch size** = 4, 8, 16

Chosen hyperparameter combinations

Table 10 lists the chosen hyperparameters of the two best-performing models, for the task of emotion and empathy classification respectively.

Model	Classification task	Fine-tunings	Learning rate	Adam eps.	Grad. acc. steps	Batch size	Epoch with best accuracy
RoBERTa	Emotion	2	1.35×10^{-4}	1×10^{-8}	1	16	10
T5	Empathy	1	1×10^{-4}	1×10^{-8}	2	8	16

Table 10: Selected hyperparameters of the two classifiers used in our implementation.

E Questionnaire responses

Tables 11 and 12 summarise the responses given by trial participants to the multiple-choice questions in the evaluation questionnaire.

Statement	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
The chatbot was good at guessing my emotion	0/16	2/16	4/16	7/16	3/16
When interacting with the persona named Kai, I found that they displayed empathy in their responses throughout the conversation.	0/16	0/16	2/16	12/16	2/16
When interacting with the other personas (any apart from Kai), I found that they displayed empathy in their responses throughout the conversation.	0/16	0/16	4/16	9/16	3/16
When interacting with the persona named Kai, I found the conversation to be engaging.	0/16	1/16	4/16	10/16	1/16
When interacting with the other personas (any apart from Kai), I found the conversation to be engaging.	0/16	1/16	4/16	9/16	2/16
Overall, the platform was useful.	0/13	0/13	1/13	10/13	2/13

Table 11: Proportions of responses by level of agreement/disagreement with the statements presented in the questionnaire.

Robert	Gabrielle	Arman	Olivia
8/27	4/27	8/27	7/27

Table 12: Number of reported chatbot interactions by chosen persona (excluding Kai).

Below we list the trial participants' open text feedback.

Please describe the overall emotional impact of your experience.

- Well it was ok. There were more interactions this time but again limiting the feelings to just 4 or 5 categories doesn't seem helpful. Why "being tired" is considered as "sad" emotion. No need to push people to sadness. Instead the chatbot can use "tired" in their response and say "I understand you had a hard day. I can help you by suggesting you doing one of these exercises, you will gain lots of positive energy for your rest of day." Or if it's night time: "I understand you had a hard day. I can help you by suggesting you doing one of these exercises, you will have a good night sleep." Isn't it better this way? By the way, are you checking the local time of the user in order to see if it's day time or night? For the above example, it seems it's better to be checked :)

- How come I hadn't worked for sometimes, but after I practiced the protocol, I easily could change my mood to a better level and I think that's because of the profound impression of the protocols the time we worked on. Also this new platform makes this good impression double as you are interacting with someone to do the process and it's close to the real conversation.

- Again, as last time, I think this question needs to be rephrased and made more specific? (what do you mean by 'emotional impact'? Are you asking about the experience of using the chatbot or the experience of doing the protocols, or a mixture? ...) Trying to decode the question based on my own interpretation, I would like to say the following: The chatbot was considerably more organic and empathetic in conversation and did make me feel heard and supported to a reasonable extent. Regardless of the chatbot, the experiential cycle of reflecting on my current emotional state, practicing a SAT protocol, and re-evaluating my state of mind had a positive impact on my focus and well being.

- It was overall a very emotional experience. I could connect with Kai and Robert and Gabrielle.

- The suggestions was appropriate and they helped me to improve my emotions.

- The experience is quite powerful, but due to the SAT protocols. The bot helps, but I feel it is quite limited, the responses are often repeating.

- I like the fact that the bot reminds me of the protocols.

- I enjoyed some protocols more than others. Exercise 13 and 20 were personal favourites as I felt as though they really boosted my mood.

- With the conversation, I can gradually adjusted my emotions and let myself feel better.

- I feel content after the protocols.

- Only some impact. Definitely limited, but detectable.
- It is very useful and great.
- It was good.

Do you have any further suggestions for how the platform can be improved?

- I'm not sure if you have tried different devices to try and see how this chatbot page works or not. I had difficulty choosing the bot's responses and suggestions, scrolling down the page, reading the chat and the instructions on both iPhone and iPad using Safari. I tried Chrome on my iPhone which was slightly better as I could see the chatbot messages under the text box but again I couldn't scroll down the instructions part to see the protocols descriptions. I could only the see first page of it which ends with this sentence: ”Please scroll down ...”. Furthermore the last time I tried the chatbot, I was trying to scroll up to see what were the previous protocol suggestions then the whole chat got restarted! So I was like well, I think I'm done trying to work with this chatbot.
- As I mentioned, the language and flow of questions and responses was really empathetic. You are definitely on the right track on this. There are a couple of details, I would like to comment on below. The acknowledgement of our feelings was very effectively done, but I feel in conversations with multiple follow-up questions (typically when discussing negative emotions), it was a little overdone, which made it less organic than a similar conversation with a real person. In real life, when we acknowledge someone's feeling, we wouldn't typically repeat it verbally at the beginning of every other sentence, but would use more subtle ways to expand on the safe space created by the original acknowledgement (which could involve, more listening and less interruption - even if the interruption has an empathetic language). Of course, a great part of this is communicated non-verbally, which is not possible with a text chat bot, but I think there is still a lot that can be done to show more subtle manifestations of compassion even in text. One thing that I found slightly counter-productive was a specific way of ending the discussion on negative feelings. After a protocol was suggested and practiced, we would be asked if it helped and several times, I noticed that this question was immediately followed by something like ”I hope this helped” or ”I hope you feel better”. In real-life empathy, this is not a good practice, as we're practically imposing a bias towards ”feeling better”. If we try to help someone in distress, and give them a chance to report back, the correct follow-up would be to ask an open question (how are you feeling now? how did it go?) followed by empathetic listening (in the case of the chatbot, perhaps it would be best to use the same AI to again interpret the response and follow up based on that).
- I think the evaluation of how helpful each session was should be more nuanced. The two choices ”I feel better” and ”I feel worse” were often inadequate for me to describe the outcome. I suggest a Likert-scale question at the least, and even

better, a series of Likert-scale statements for this purpose. The current response recognition seems to mostly favor more pronounced/extreme feelings. In response to my answers which were often something like, 'I feel slightly worried' or 'I am ok, but would love to boost my motivation', the chatbot would generalize my feelings to their more extreme versions (e.g. "nervous/scared"). While the recommended protocols would probably remain the same, I think it would be helpful to add this nuance to the way these feelings are recognized and acknowledged. This is, in my view, particularly important for more advanced use of SAT for non-clinical users for whom the tool can be used to enhance mental and emotional well-being rather than to simply alleviate distress in its more evident forms.

- I think it was great, maybe some more empathetic phrases can help individuals in trouble.
- Emotion recognition can be significantly improved with a corpus of synonyms. Responses could be more varied and flavorful. Often silly, funny even ridiculous responses can help.
- The persona must ask deeper more details questions before suggesting protocols. Regardless of how much of the users' answers are actually used, the impression that someone is listening to me at the other end of the line, is very important I think. The responses of the persona will be better received if they are in audio/video format, either through animation or ideally AI assisted recorded video.
- I think it would be a great idea if the bot itself gave a short reminder of the protocol item that it suggests. Also I think the bot tries to show too much empathy in an artificial way. Also, I think the bot is unable to understand complicated emotions. It simplifies the emotions a little too much.
- Potentially introduce more of a conversation rather than just clicking buttons to answer questions, but of course I understand this might be hard to implement.
- One of the versions of questions asking for more information gave me the impression that I would have to type more information instead of selecting, so I chose not to provide this information (led to worse suggestions, the protocols I received did not appear to be relevant). Not sure if it would be better to just ask the questions instead of asking for user's approval on this - I noticed that the chatbot asks the user a lot more than in the previous version, which is good so the chatbot has permission to ask these questions but also leads to potentially worse suggestions. I noticed 4 protocols were provided as suggestions each time as opposed to fewer suggestions in previous versions depending on the case, though some appear to have been randomly added and did not appear to be appropriate. Instead of asking the user each time for their name, allow them to enter this in a separate page (or save this the first time the user asks for it). For this trial, this could have been manually filled in to avoid asking for it each time. However, integrating this with sentences to personalise the

conversation was effective. It would be nice if it was easier to switch between personas without needing to refresh e.g. by adding a button that allows you to select from the options, which then restarts the conversation.

- I'm not sure what is the purpose of the panel (Protocol viewer) on the right-hand side? It was empty all the time and was taking up a half of the screen. Is that intentional? Do I have to feel either better or worse after the session? One could feel the same, but that answer was not presented.
- It would be more engaging if the actual personas take you through the protocols themselves, step by step. This could make the chat more interactive. Just a thought.
- UI suggestions: Once you pick a persona, that persona's avatar should be the only one shown (preferably to the left of the text, just like iPhone). Right now, you still see all the other faces below and that's not very immersive to the one persona. Second, the text parts of the SAT protocol should be just integrated into the chatbot interface itself, maybe even with images. The bot asking you to read something somewhere else (even if it's in the panel right next to you) is not as engaging.
- I think if the bot can use some emotional pictures for guessing emotions, it would be more comfortable.
- It appears as if the responses are hardcoded. I suggest its answers should be dynamic. I asked a question like 'how are you?' and the answer was standard.

Please provide any additional comments you have.

- Sorry that my comments seems discouraging. I totally understand how much effort you and the previous team have put to make this work better and I really appreciate your hard work. Please don't get me wrong. I'm sure by creating and improving this chatbot, you all are going to make a huge difference in lots of people's lives. Thanks for the hard work! :)
- Overall, this was a huge step in the right direction. I look forward to seeing the later versions as well. A few more comments and suggestions: The interface has a lot of room for improvement. The current version is not compatible with all devices. There are also several details that can be improved, which is beyond the scope of this feedback. I suggest you collect user feedback specifically on this aspect later in the research. I mentioned this in my feedback to the previous version (and there was a good attempt at improvement in this version, but my feedback still stands): Asking the participant to "go, do the protocol, and come back" somewhat disrupts the organic flow of the mentor/mentee conversation. Ideally, I would like to be guided by the chatbot persona in doing the protocol. Aspects of this was incorporated in the questioning (which I really appreciated), but I think there is still room for making it more streamlined. One feature of the last version, which for some reason was removed in

this version, was the option to choose a different protocol than those recommended by the chatbot. I think this feature should not only be kept, but also improved to help the AI learn about why respondents are choosing those protocols and how the knowledge can be used to have better recommendations for them in future sessions. The addition of personas was great. I think more work is worth doing in making the persona relatable and ”real” (for the first step, the chosen persona's avatar can be shown throughout the conversation. Right now, we'd choose a persona, but still see the face of everyone else along with our chosen identity.)

- The responses feel quite general. This in turn might make the user sceptical towards the effectiveness of the suggested protocols, even before trying them.

- I think it is better to focus on improving the bot as a tool to teach us the protocol itself at a deeper level.

- The AI understood the main buzz words of my emotions. It did get confused once when I said ”I was feeling so so” and it interpreted it as I was feeling happy, but in general it understood my emotions.

- The response from the chat bot maybe should be displayed one by one to make it like a real person?

- Good use of dialogue, particularly for users with negative emotions. When speaking with Kai, giving a negative emotion and saying I feel better, the follow-up suggestions did not seem to be consistent with the suggestions before (perhaps some were randomly shuffled?). Good use of statements as questions to mix the dialogue up a bit - more varied dialogue. For Arman, the following dialogue was vague: ”Is there anything going on in your personal life that is causing you to feel this way?” - the wording is implicitly linked to friends/family but this could mean anything. When I said yes to this question, Protocol 1 was one of the suggestions - seemed inappropriate. For Olivia: ”I'm in tune with your emotion and we will understand how to overcome it.” - seems a little awkward to say, less realistic. ”Have you felt or shared any of these feelings with anyone in your life:” ¡- perhaps could include some more details about how this would link to the feeling the person has. Protocol 5 was one of the suggestions when I again selected yes to the personal life question - did not seem appropriate. Overall, some of the prompts were inappropriate but the dialogue was largely empathetic and at times the different personas were engaging. It's great to see how this has developed over time.

- I think you are relying too much on self motivation to follow through on all the protocols. For someone depressed for example, not sure there's motivation to (go do X and Y, and come back here and tell me if it worked). If the bot appears like it doesn't know the details of what it's asking you to do, it's not very reassuring.

- Great well done.

Below we list the clinician's open text feedback.

Please describe the overall emotional impact of your experience.

I was able to engage with the issues raised examine my own socio-political views as well as the impact of some of my childhood events on my attitude and behavioural responses.

Do you have any further suggestions for how the platform can be improved?

1. I found the characters' responses identical, regardless of their names and appearances. 2. My emotions were correctly guessed but they were limited to a few. 3. Empathic responses were 'robotic' and stereotypic regardless of the emotional expressions. 4. I did not detect any difference between the characters' empathic responses.

Please provide any additional comments you have.

Overall, the programme provides the opportunity to engage in SAT.