This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test

Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan

Abstract—Mental disorders are widespread in countries all over the world. Nevertheless, there is a global shortage in human resources delivering mental health services. Leaving people with mental disorders untreated may increase suicide attempts and mortality. To address this matter of limited resources, conversational agents have gained momentum in the last years. In this work, we introduce SERMO, a mobile application with integrated chatbot that implements methods from cognitive behaviour therapy (CBT) to support mentally ill people in regulating emotions and dealing with thoughts and feelings. SERMO asks the user on a daily basis on events that occurred and on emotions. It determines automatically the basic emotion of a user from the natural language input using natural language processing and a lexicon-based approach. Depending on the emotion, an appropriate measurement such as activities or mindfulness exercises are suggested by SERMO. Additional functionalities are an emotion diary, a list of pleasant activities, mindfulness exercises and information on emotions and CBT in general. User experience was studied with 21 participants using the User Experience Questionnaire (UEQ). Findings show that efficiency, perspicuity and attractiveness are considered as good. The scales describing hedonic quality (stimulation and novelty), i.e., fun of use, show neutral evaluations.

Index Terms—Conversational user interface, natural language processing, sentiment analysis, mental health, mHealth.

1 Introduction

Mental disorders affect 29% of the global population in their lives [1]. Every year, 25% of adults and 10% of children are affected [2]. The most common mental disorders are depressive disorders and anxiety disorders. In 2017, 322 million people suffered from depressive disorders and 264 million from anxiety disorders worldwide [3]. Apart from the fact that mental disorders impact on people's quality of life, they are one of the most common causes of occupational disability [4] leading to high economic costs. In Switzerland, it is estimated that more than CHF 11 billion - including indirect costs that are associated with untreated mental disorders - are spent per year, e.g. for reduced labour productivity and suicide complications [5].

Mental disorders are usually treated by pharmacotherapy or psychotherapy [6]. However, there is a global shortage of human resources for delivering such mental health services. In developed countries there are nine psychiatrists per 100,000 people available [7], while in developing countries there is one psychiatrist per ten million people [8]. According to the WHO, about 45% of people in developed countries and 15% of people in developing countries have access to psychiatric services [9]. Leaving people with mental disorders untreated can increase suicide attempts and mortality [10]. There is a need for providing support and self-help between therapeutic sessions and for patients waiting for treatment by mental health service providers. To address this problem, conversational agents have arisen interest in recent years, particularly in psychoeducation, behaviour change and self-help.

Conversational agents or chatbots are text-based dialogue systems integrated in mobile apps or web pages. The chatbot simulates a realistic conversation partner by giving the user appropriate written answers in a language that he or she understands. Chatbots were mainly used in marketing to enhance customer experiences. A recent review of Laranjo et al. confirms that the use of conversational agents with unconstrained natural language input capabilities for health-related purposes is an emerging field of research [11]. Vaidyam et al. concluded that the mental health field could use conversational agents in psychiatric treatment [12]. According to a systematic review [13], there are currently 41 different chatbots for mental health reported by 53 studies. About 43% of those chatbots were implemented in the United States of America. They focused on different use cases including therapy (e.g. Woebot), training (e.g. LISSA), and screening (e.g. SimSensei), and mainly on users suffering from depression or autism. So far, there is only limited clinical evidence that mental health chatbots are reducing symptoms. Woebot was tested with students with depression [14]. The students who used Woebot significantly reduced their symptoms of depression over the study period as measured by the depression questionnaire PHQ-9 while those in the information control group did not. The control group had to read a self-help book.

Many existing mental health chatbots restrict the conversation to a limited set of patterns in form of predefined answers. This limits the expressiveness of interactions for the user. Furthermore, currently only mental health chatbots in English are available.

In this paper, we are focusing on conversational agents with unconstrained natural language input in German. More specifically, the aim of this work is to develop a mobile application with integrated chatbot that supports the user in regulating his or her emotions. Emotions will be recognized directly from the conversation which requires methods for analysing free text input with respect to emotions. The following questions serve as a basis for the development

- *Bern University of Applied Sciences, Bern, Switzerland .*
  E-mail: kerstin.denecke@bfh.ch

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

and implementation of the application:

- How can emotions be recognized in chatbot conversations?
- How can detected emotions be used to support a user in regulating his / her emotions?

The paper is structured as follows. Section 2 describes the background of cognitive behaviour therapy (CBT) and introduces the state of the art in conversational agents for mental health. Furthermore, an overview on emotion analysis from natural language text is given. In Section 3, we describe the requirement analysis process, the system development and used frameworks. The mobile application including the chatbot is introduced in Section 4. A usability test was performed with methods and results summarized in Section 5. The paper finishes with discussions (Section 6) and conclusions (Section 7).

2 Background

2.1 CBT and emotion regulation

We grounded the development of the application, especially the information basis of the chatbot, on knowledge and best practices of CBT. CBT was initially developed for the treatment of mild to moderate depression [15]. However, it is also in use for treating other mental disorders such as anxiety disorders, panic disorders, bipolar disorders and post-traumatic disorders. The basic assumption of CBT is that psychiatric disorders arise and are maintained due to distorted cognition (thoughts and attitudes). In modern CBT, the treatment pays increasingly more attention to the emotional aspects since emotions and their regulation impact on mental health [16].

Basically, emotions are considered subjective perceptions that persist over a short period of time and relate to specific events, persons or objects. According to Richard Graph's C-I-E (Cognition, Intuition, Emotion) theory, there are seven basic emotions: Fear, disgust, anger, joy, grief, guilt and shame [17]. Finally, emotion regulation concerns influencing the type, intensity and duration of emotions into a certain direction.

2.2 Conversational agents in mental health

Existing studies and reviews show that mobile apps with integrated CBT can be successfully used for the treatment of psychiatric disorders. In a review by the Bolton University, the effectiveness of mHealth applications with CBT was investigated. Half of the studies focused on the treatment of depression with generally positive results [18]. A marginalized control study was conducted with 300 participants using the mobile Web app MoodHacker [19]. An online self-assessment survey was realized at the beginning and after six weeks. During the six-week follow-up, significant effects of depressive syndromes, behavioural activation, negative thinking, knowledge, work productivity, absenteeism and disability were observed.

Clinical outcomes for the use of mental health chatbots are still rare as a recent systematic review by Abd-alrazaq et al. [13] shows. The authors identified 41 different chatbots in mental health, mainly implemented as rule-based and stand-alone software, e.g. Wysa and WoeBot. In contrast to other digital interventions in mental health, chatbots aim to increase the adherence to the intervention [14] [20] [21]. Chatbots process the user input, and offer responsive, guided conversations and advice to help users in current mental health challenges. The bots normally ask a user on a daily basis on his emotions, thoughts, and behaviour. Some systems passively track users' movements via the accelerometer integrated in the phone [22].

The chatbot Wysa [22] provides a mood tracker and can detect negative moods. If necessary, it suggests a depression test and recommends seeking professional help. To support the relief of anxiety, depression and stress, there are mindfulness meditation exercises integrated in the app. The chatbot was tested in a study with a total of 129 participants [22]. The participants were divided into two groups (frequent and occasional users). The quantitative results show that frequent users had a higher, average improvement in their mood than the group of occasional users. Two thirds of the users perceived the app as positive. They replied that the conversation with Wysa was helpful and stimulating.

Woebot is a chatbot that uses CBT strategies to help users cope with symptoms of anxiety and depression [14]. The chatbot allows to enter emotions by selecting terms from a list of suggestions. This limits the user to comprehensively express his or her current emotions and feelings. In current publications, it is not mentioned on which psychological evidences the system is based on. In a study comparing people who interacted with Woebot versus a group of people who read a self-help book 12 times over two weeks , those who used Woebot had a reduction in their symptoms. Another chatbot, Replica[1], allows users to reply in their own words to chatbot comments, but the chatbot does not understand the context and therefore gives inappropriate answers or changes the subject.

Mental health apps are easy accessible and easy to use. They can be consulted whenever users feel sad, anxious, stressed, or just want a distraction. They are also significantly less costly than face-to-face interventions such as CBT [21]. Vaidyam et al. found out that chatbots showed potential to support psychoeducation and self-adherence [12]. Users are satisfied by interacting with such systems, indicating that they could provide an extension to psychiatric treatment. Limitations of the existing chatbots are that they are only available in English and the chatbots are asking for emotions, but are normally incapable of determining emotions based on natural language user input. In this paper, we introduce SERMO, a chatbot with integrated CBT interventions in German enabling unrestricted natural language user input. It differs from the available systems by integrating natural language processing (NLP) and emotion analysis methods in order to automatically determine emotions from the user input. In contrast to existing decision-tree based systems, our system does not rely on strict patterns, but on syntactic and semantic similarities between user input and stored expressions. Furthermore, with this paper we address the issue that existing mental health chatbots are rarely described in sufficient detail [13]. We describe details on the underlying psychological evidences, technical implementa-

[1] https://play.google.com/store/apps/details?id=ai.replika.app

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

tion and dialogue structure. This will allow researchers and practitioners to judge the quality of the underlying evidence base.

2.3 Emotions in chatbots and natural language

Emotions can be detected in text [23], voice [24] and faces [25] with varying reliability. In general, emotion recognition is a two-step procedure which involves extraction of significant features and classification. This general principle holds true for all three sources of emotion detection, text, voice or faces, but the relevant features differ. Feature extraction determines a set of independent attributes, which in sum can characterize an expression of an emotion. Features for emotion recognition from faces include for example specific distances or angles in the face determined from recorded images (e.g. angle of eyebrows [26]). For classification in emotion recognition the features are mapped to one of various emotion classes like anger, joy, sadness, disgust, surprise, etc. The feature attributes and the chosen classifier impact on the classification quality. The classification is often challenging since multiple emotions can be expressed at the same time [27].

We are focusing on emotions expressed in natural language. There are two basic procedures: lexicon-based or machine learning-based methods for analysing emotions in text. A lexicon-based method uses an emotion term lexicon that for each emotion contains terms that could be used to express this particular emotion. Through lexicon lookup, the input sentence is matched with the lexicon terms. The matches are aggregated to determine an emotion. A problem with this method is, that the context remains unconsidered. This is sometimes important for a correct emotion classification since a term can change meaning in different contexts. In contrast, machine learning-based approaches are based upon labeled training data and could consider the context. Prominent examples for algorithms are Naive Bayes and Support Vector Machines (e.g. [23]).

Analyzing emotions or sentiments resulting from interactions with chatbots has so far only rarely been addressed. There are multiple ways to enable a chatbot to choose an emotion category for a response. On the one hand, the chatbot can be equipped with a personality and background knowledge. On the other hand, training data can be used to find the most frequent response emotion category for an emotion in a given response and use this as the response emotion. Previous research by Skowron proposed affect listeners, i.e. conversational systems that can respond to users' utterances on a content-, but also on an affect-level [28]. Zhou et al. [29] describe an emotional chatting machine that can generate appropriate responses fitting in content and emotion to a users' response. The architecture consists of a recurrent neural network enabled with GRU cells with attention mechanism. It contains three different mechanisms for generating responses with a specific emotion: External knowledge serves to model emotions explicitly using an external emotion vocabulary. Internal memory captures emotion dynamics and finally, different emotion categories are represented as embedded vector. Socher et al. introduced a sentiment treebank that includes fine-grained sentiment labels to parse trees of sentences. On this treebank, they applied recursive deep models to predict sentence level sentiment. With a complicated treebank annotation, the proposed method has recognized the negated sentiment in a better way and achieved more than 80% overall accuracy [30].

Sentiment and emotion analysis in a medical context has been mainly addressed for web content. Denecke and Deng reviewed the state of the art and studied the challenges of sentiment analysis in medical settings [31]. They found out that given the varying usage and meanings of terms, sentiment analysis from medical documents requires a domain-specific sentiment source and complementary context-dependent features to be able to correctly interpret the implicit sentiment. The challenges of sentiment and emotion analysis in mental health chatbots have not yet been considered so far. Furthermore, health applications equipped with emotion and sentiment analysis are still missing.

Although there are limitations of lexicon-based approaches, we decided on the current implementation to integrate only a lexicon-based approach into SERMO. The main reason is that training data in German is missing, while emotion lexicons for some emotions are already available. SERMO integrates a method to analyse a user statement to select an appropriate, motivating or encouraging response when given a specific user emotion. The system is implemented in a way, that in future, the system could create in parallel an annotated data set: Emotions recognized by SERMO have to be confirmed by the user which could be used to collect labeled data.

3 Methods

In this section, the methods for collecting requirements and developing the application SERMO are described. Methods for testing the usability of the application are introduced in section 5 along with the test results.

3.1 Requirement analysis

The application requirements were determined by means of a literature search, interviews and discussions with experts. More specifically, four psychologists of different clinics in Switzerland and Germany were interviewed. The interviews focused on the current treatment of mental disorders and on current practices to accompany patients in the time between two therapeutic sessions. The collected requirements formed the framework of the implementation. The literature search focused on mental diseases in general, psychotherapy, mental health applications with conversational user interfaces, emotion recognition in free text, and sentiment analysis. Results were retrieved from Pubmed and Google Scholar.

Furthermore, persons suffering from mental diseases were asked for app functionalities that would be of help to deal with their disease. Four young people aged 16-25 with diagnosed depressions and one 59-year-old man with bipolar disorder were interviewed. The patients had been suggested by the psychologists considering their current mental state.

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

3.2 System development

The chatbot was developed using the Syn.Bot framework (https://www.nuget.org/packages/Syn.Bot/). It contains OSCOVA (https://oscova.com) and an official SIML (Synthetic Intelligence Markup Language) interpreter. The framework is platform independent. We use OSCOVA to realize the chatbot. Compared to other chatbot frameworks, OSCOVA does not have a hybrid decision tree or does not rely on strict patterns. Instead, it relies on the semantic and syntactic similarities between user input and stored expressions. OSCOVA also allows developers to use machine learning and NLP functions. Another advantage is that OSCOVA does not require a connection to the online API and can therefore be used in an offline setting.

OSCOVA consists of five different components: expressions, entities, contexts, intents, and dialogues. An *expression* is a pattern that defines user input. The expression attribute is used to decorate an intent method by triggering user input expressions. *Entities* are pieces of information within a user message that a developer would like to extract. *Entities* are associated with an entity type like "dateTime" and "emotion". The *context* represents the current context of the conversation, i.e. the conversation state of a user session. An *intent* is any action that the bot is supposed to execute when the user message is similar to an *expression*. A *dialog* in OSCOVA is used to group together a collection of related intents and actions. Dialogues determine which responses must be returned to the chatbot's user input.

The application was developed with Xamarin.Forms (https://docs.microsoft.com/en-us/xamarin/xamarin-forms/). Xamarin.Forms was chosen because the OSCOVA framework exists as .NET NuGet to facilitate referencing OSCOVA in .NET projects. An advantage is that Xamarin.Forms provides a cross-platform interface toolkit for .NET developers. Large parts of the development results can already be used for all platform implementations. We developed a native Android platform application with Xamarin.Forms. The underlying database uses SQLite.

4 SERMO - System Overview

In the following, we are going to describe the scenario underlying the system development. It was developed based on the collected requirements that are also summarized. Afterwards, the system architecture and functionalities are introduced. Finally, technological details on the implemented chatbot conversation and emotion analysis algorithm are provided.

4.1 Requirements

From the expert and patient interviews, we decided on the following scenario: Mona is a 25-years old student. During the semester, she has to pass many exams and projects to hand in. She is under stress and constantly has negative thoughts and emotions regarding her ability to complete the studies successfully. She reproaches herself and blames herself for everything. One day, she collapses at school. As a consequence, she starts psychotherapy and sees a psychotherapist once a week for an hour who exploits CBT methods. In addition to the sessions, Mona is asked to document her emotions and to deal consciously with her problems. For this purpose, she desires an app that supports her in keeping a diary of thoughts and emotions and that supports in coping with her mental health problems between therapy sessions.

The functional requirements can be grouped into six categories. They concern the login, the chatbot, diary, list of activities, information provision and notification. It should be possible to log in to the app using a code. The diary should allow to enter a daily goal, to record the daily mood and to document an event. The event is stored according to the ABC schema (situation, thoughts, emotions) [32]. Depending on the users' emotion, the system should suggest activities or exercises to the user and should contact the user at least daily. Further, the user should be enabled to access diary entries and to get an overview on the mood development over the past month. Desired chatbot functionalities include:

- User can frame answers in his own words.
- User can select predefined answers.
- Chatbot recognizes emotions in natural language user input.
- Chatbot suggests activities and exercises for regulating emotions.
- Chatbot creates an entry for an event.
- Chabot stores the mood of the user on a daily basis.
- Chatbot stores specified goals of a user.
- Chatbot reminds user on appointments.

Non-functional requirements include that the system should run on Android 7.1 or iOS 12.2.

4.2 Architecture

Figure 1 shows the system architecture of SERMO. The collected data on events, daily mood etc. are stored in a structured manner in an SQLite database in the internal storage of the mobile phone. User input from the chat is processed using the OSCOVA interpreter. It includes an NLP component that realises the emotion recognition. The OSCOVA interpreter determines the context and intentions. The NLP component, in particular the emotion recognition method, exploits a knowledge base which is a lexical resource with lists of words that express the emotions that can currently be detected. More details on the emotion recognition algorithm are provided in section 4.4

4.3 Functionalities

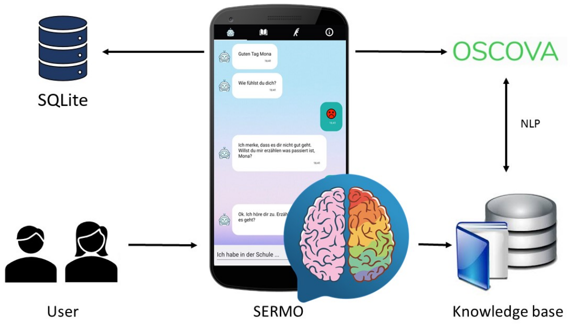
Following the collected requirements, SERMO provides four main functionalities: 1) interaction with the chatbot, 2) provision of activities and exercises to train the attentiveness, 3) diary of events with associated emotions and 4) information provision.

4.3.1 Chatbot

In interaction with the user, the chatbot asks for the current mood, runs an ABC dialogue to retrieve information on a current event that impacted on the user as well as the emotion associated with the event. Based on this, it suggests suited activities and exercises (see Fig. 2). The content of the

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing



Guten Tag Mona 16:41

Wie fühlst du dich? 16:41

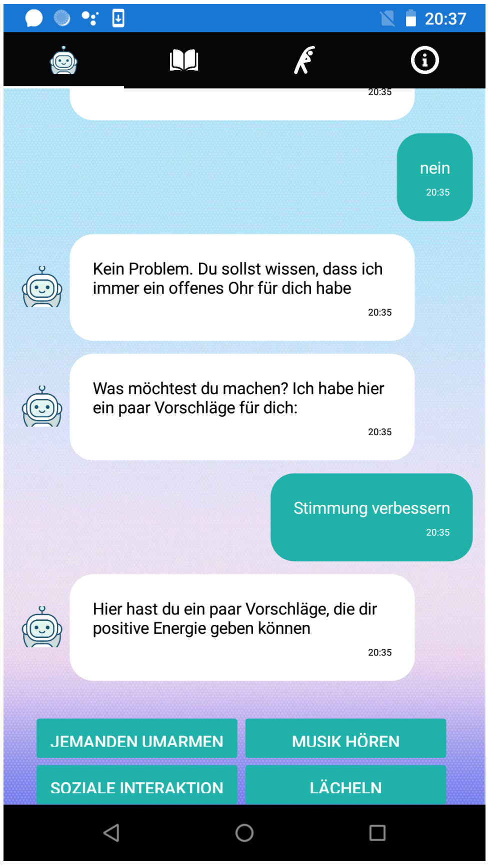
16:41

Ich merke, dass es dir nicht gut geht. Willst du mir erzählen was passiert ist, Mona? 16:41

Ok. Ich höre dir zu. Erzäh... es geht?

Ich habe in der Schule ...

Fig. 1. System architecture of SERMO. The knowledge base contains the chatbot brain, i.e. information on how to react to an emotion. Further, a list of emotion terms is captured in the knowledge base as a basis for the emotion analysis algorithm. The SQLite database is used to store diary entries.



20:35

nein 20:35

Kein Problem. Du sollst wissen, dass ich immer ein offenes Ohr für dich habe 20:35

Was möchtest du machen? Ich habe hier ein paar Vorschläge für dich: 20:35

Stimmung verbessern 20:35

Hier hast du ein paar Vorschläge, die dir positive Energie geben können 20:35

Fig. 2. Screenshot of the chat: SERMO asks whether the user would like to get some suggestions on how to improve the mood. It suggests hugging someone, listening to music, social interaction, smiling. Once the user selects an activity, SERMO will explain, why this activity is helpful for changing the mood.

chatbot was grounded on scientific literature, expert input and assessment of existing apps. SERMO integrates four contexts (where a context is a topic of a specific part of a conversation): daily mood, emotion recognition, supporting measurements and information / other activities.

Patients with depression or bipolar disorder often suffer additionally from mood disorders. Keeping an emotion diary can help people monitor and control their emotions [33]. Previous research showed that mobile mood applications helped users to experience patterns in their mood, to improve their mood and to cope better with stress and emotions [34]. Therefore, SERMO asks daily the user about his mood. It can be selected on a slider scale from one to five. When the mood is rather positive, tasks are suggested to train the strengths and resources of the user. In case the mood is rather negative, the chatbot asks for the triggering event of the mood. For this purpose, the ABC theory has been implemented in the chatbot.

The ABC theory was developed by Albert Ellis. It follows the approach that consciously or unconsciously perceived stimuli are evaluated and these evaluations lead to certain feelings and behaviors [32]. The Activating Event (A) represents an external or internal event or situation. The Belief (B) comprises attitudes and thoughts regarding the event and Consequences (C) reflect feelings and behaviors. SERMO collects in a dialogue with the user information on A, B, and C and exploits the user input to determine the emotion. The emotion analysis (see section 4.4) currently distinguishes five emotions: fear, anger, sadness, joy, grief. In case the user decides to regulate the emotion, SERMO first offers information on the detected emotion and then suggests exercises or activities. For example, if the emotion anger is detected, SERMO asks whether it is a justified anger or not. If justified, the user is asked whether he or she would like to change the situation. If unjustified , the chatbot tries to distract the user by asking positive questions such as ”What are you proud of?”, ”What do you like doing in your spare time?”.

For realizing the chatbot conversation, 13 dialogues have been developed with OSCOVA. They are described in table 1. They cover the various interactions triggered by an emotion or mood expressed by the user. An example of the flow of trigger events is shown in figure 3.

4.3.2 *Activities and exercises*

Several studies show that certain activities have a positive impact on patients of various disorder groups [35]. To reflect this aspect, SERMO provides users with a list of pleasant activities divided into four categories: Mindfulness exercises, relaxation exercises, leisure activities and others that are accessible also outside the chat conversation. The list was adopted by a psychologist. A list of pleasant activities is helpful for users who are unable to develop sufficient ideas for positive activities. The user can mark activities as favorites within the app. After performing an activity, the user can enter his experiences and emotions associated with the activity in his SERMO diary.

Mindfulness exercises are a form of meditation. They are increasingly used within CBT. In this context, mindfulness means to focus one’s attention intentionally and non-judging on the conscious experience of the present

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

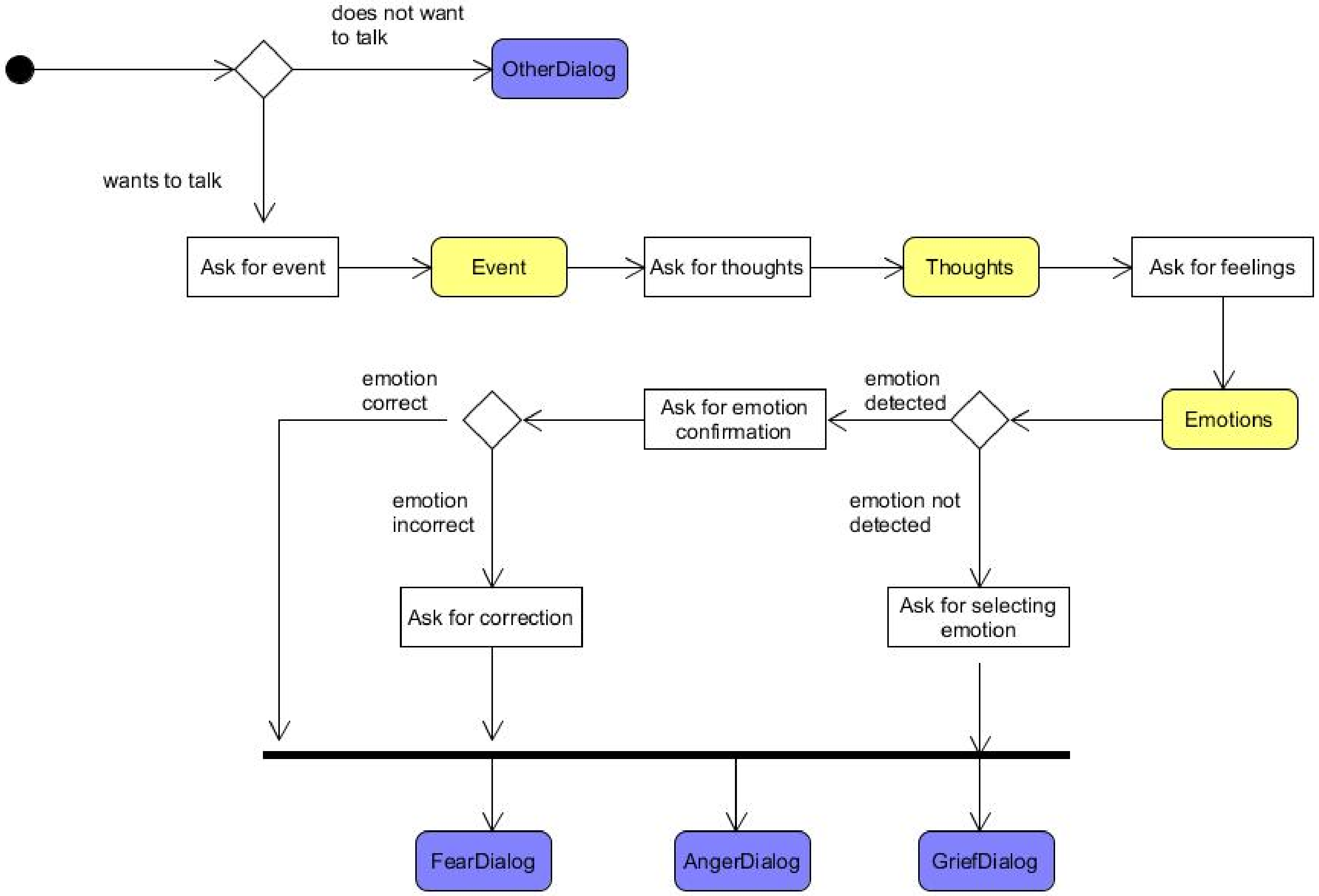


Fig. 3. Flow chart of the Emotion Dialog. Following the ABC theory, the user is asked about an event, his thoughts and feelings. SERMO then determines the emotion. If recognized, SERMO asks for confirmation. If recognized incorrectly, the user can enter the correct emotion. Finally, SERMO proceeds by handling the correct emotion (FearDialogue, AngerDialogue, GriefDialogue)
.

moment [36]. In the application, the mindfulness exercises breathing, sitting meditation, body scan and mindful yoga are suggested and integrated as audio.

4.3.3 Diary

In the diary, the user can record his mood on a daily basis (see figure 4). The mood is recorded using a slider with five different smileys ranging from good to bad. It is possible to record several moods on one day. In this case, the average mood of the day is displayed in the day view. The diary contains a day and a month view. In the day view, the user can quickly and easily document his mood and the things he has experienced. The month view shows the course of the average mood for a period of a month. In this way, the user can observe changes in his mood over a longer period of time.

In CBT, psychotherapists work with goals. Together with the patients, goals are defined for a short time and the patient tries to achieve this goal. In SERMO, patients can enter the discussed goal. With the diary and set goals, a user can understand himself in a better way, observe the course of illness and if necessary, based on the aggregated view shown in the app, discuss it with a therapist.

4.3.4 Information

The information section of SERMO comprises four topics: SERMO, CBT, emotions and advice. In the tab “SERMO”, the app is explained with all available functions. The basics of CBT are introduced under tab “CBT” and the seven basic emotions (fear, disgust, anger, joy, grief, guilt, shame) are explained in tab “Emotions”. Emotion regulation primarily concerns supporting a user in identifying and naming his feelings, emotions and thoughts. In order to achieve this goal, it is important that the user knows what emotions are and what they mean. This is not always as self-evident as thought and is part of emotion-focused therapy. SERMO therefore explains the basic emotions in the information section. Beyond, addresses of professional counselling centres from all over Switzerland are provided. The information on the counselling services is relevant, as mental illnesses should not be underestimated and must therefore be treated by specialists as early as possible. The app is not expected to replace professional help, but to encourage users for self-help and equip them with appropriate information and tools.

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

TABLE 1
SERMO integrated chatbot dialogues.

Welcome Dialogue	In this dialogue, the welcoming of the user is administered. If started for the first time, the chatbot asks for the user's name and consent to the privacy policy. Then, the user is asked about his mood. If the mood was already entered up to three hours ago, the welcome dialogue is based on the last mood.
Joy Dialogue	As soon as the user states that he is in a good mood, the Joy Dialogue is called. SERMO asks for the reason of the good mood and finally asks if he wants to do a task. If so, the HashTag dialogue is started.
Normal Dialogue	This dialogue is started when the user is in a balanced mood. SERMO asks for the reason of the mood and starts the Emotion Dialogue to determine the emotion.
Sadness Dialogue	This dialogue is started when the user states he is sad. The Emotion Dialogue is started to determine the emotion.
Emotion Dialogue	After the mood has been selected, the emotion dialogue is executed. This dialogue implements the ABC theory. The user is asked about the situation or event, his thoughts and feelings. Based on the replies, the emotion is recognized and passed forward to the appropriate emotion dialogue (i.e. Fear Dialogue, Anger Dialogue, Grief Dialogue, Sadness Dialogue, Joy Dialogue).
Anger Dialogue	The dialogue handles the emotion anger. The user is informed on the different types of anger (appropriate anger and inadequate anger). Further, a pleasant activity is suggested.
Fear Dialogue	This dialogue concerns the emotion fear. The user is provided with information on reasonable and inadequate fear. Finally, he is asked to transform the fear-provoking thoughts into positive thoughts.
Grief Dialogue	The dialogue handles the emotion grief. The user is informed on the different phases of grief. Further, activities for distraction are suggested.
Other Dialogue	Further measures are proposed to the user in this dialogue. One dialogue is about improving the user's mood and the other allows the user to plan the day. In addition, the user can select mindfulness exercises or the option Nothing.
Improved Mood Dialogue	In this dialogue, various activities are suggested to the user which could improve his current mood. After having carried out an activity, he has the possibility to carry out another activity.
HashTag Dialogue	This dialogue manages specific interactions that are triggered by the user using a hashtag. In its current implementation two interactions are available: #todo show a list of tasks for the current day, #strengths shows a list of strengths of the user. Both lists can be adapted by the user.
Activity Dialogue	In this dialogue, mindfulness exercises are suggested and, if selected, the user is redirected directly to the exercise on the Activities page and the exercise is started.
Goodbye Dialogue	This dialogue manages the ending of the conversation.

4.4 Emotion analysis

The implemented emotion analysis algorithm uses a lexicon-based approach. Five emotions are recognized automatically: Fear, anger, grief, sadness, joy. The processing comprises six steps (see figure 5). First, the user input is split into sentences. Second, each sentence is tokenized. Third, stop words are removed, i.e. all words that are irrelevant for emotion classification. This includes prepositions, pronouns etc. Fourth, negations are detected, but we did not yet implemented an interpretation of negations. In principle, the meaning of emotion words with negation has to be inverted. This requires a list of antonyms for all emotion words. In the current version, the negated emotion words are excluded from further processing. Fifth, the emotion terms are determined and finally, the input is classified into one out of the five emotion categories.

The underlying emotion lexicon is the Emotional Dictionaries of SentiWS. The SentiWS is a publicly available German vocabulary for emotional analysis [37]. It covers only the five emotions listed above. It remains to the future to develop emotion term lists for the emotions guilt and shame to cover all relevant emotions. In order to deal with typos and writing errors, a fuzzy matching method is used for identifying emotion terms. In this way, words can be recognized even if they do not match 100% with words in the dictionary. A threshold value was defined for the fuzzy matching [38].

For the user input, all matches of terms with the emotion lexicon are determined. Per emotion class, the number of terms that have been identified are calculated. Finally, the user input is classified as emotion class where the largest number of terms were extracted from the input. In some cases, however, it may happen that no emotion terms are identified or there is no majority of emotion terms of one specific category. In these cases, the application responds that it could not identify the user's emotion and asks the user to select one of the five emotions. Depending on the determined emotion, the dialogue proceeds as foreseen in the emotion-specific dialogues (see table 1).

5 Usability Test

We conducted a usability test to study the user experience and quality of the app and to determine areas of improvement. Furthermore, we collected feedback from patients and experts on the app and its functionalities. The methodology and results are presented in the following.

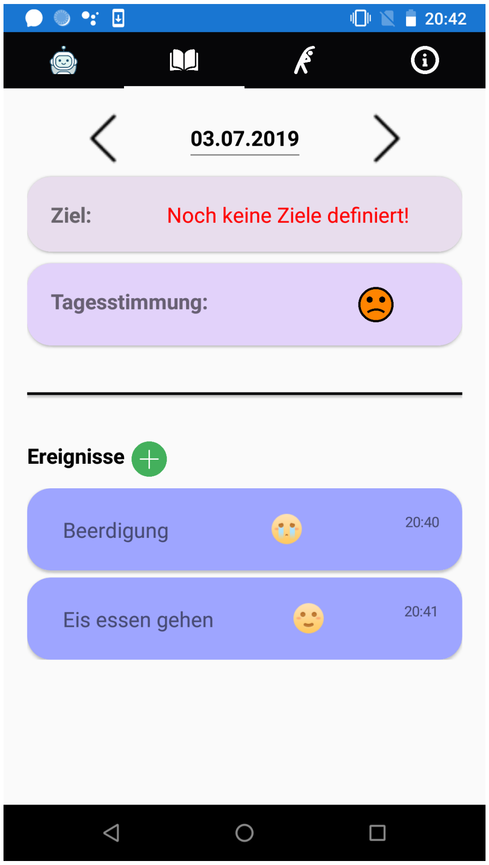
5.1 Usability test methodology

As demographic data, we collected age, gender and a personal judgment of the technical competencies on a scale of 1 (no competencies) to 10 (expert). The usability test was scenario-based comprising six tasks. The users were asked to perform the tasks to test the specific functionalities and provide feedback whether they could complete the task (yes / no) or whether and which problems occurred. The tasks included

- Define a goal,
- Enter a mood,
- Enter a current event,

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing



20:42

03.07.2019

Ziel: Noch keine Ziele definiert!

Tagesstimmung:

Ereignisse

Beerdigung 20:40

Eis essen gehen 20:41

Fig. 4. Screenshot of the diary: A goal has not yet been specified. The daily mood is rather bad (smiley). Two events had been added: One concerns a funeral; the other one a positive activity which is eating ice cream.

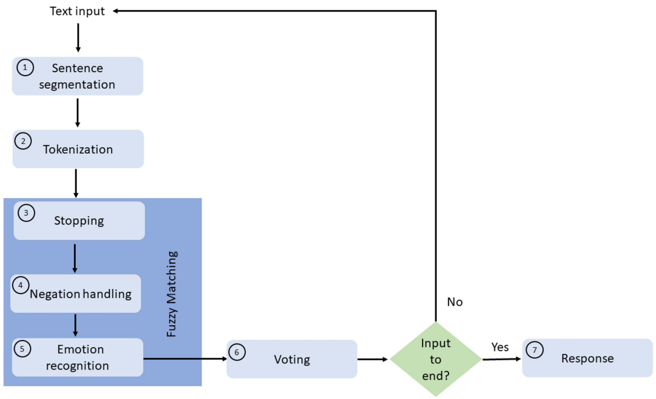


Fig. 5. Emotion analysis process

- Choose a pleasant activity,
- Run a mindfulness exercise,
- Chat with SERMO for at least 1 minute.

In the second part, the participants had to judge concrete aspects of user experience. For this purpose, we applied the user experience questionnaire (UEQ) provided by Schrepp et al. [39]. The main goal of the UEQ is a fast and direct measurement of user experience. Each item (in total 26 items) of the UEQ consists of a pair of terms with opposite meanings. Each item can be rated on a 7-point Likert scale ranging from -3 (fully agree with negative term) to +3 (fully agree with positive term). Half of the items start with the positive term, the rest with the negative term (in randomized order). The 26 items (see figure 6) are grouped into six scales: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. Attractiveness concerns the overall impression of the app. Perspicuity assesses whether it is easy to get familiar with the app. Efficiency studies whether users can solve their tasks without unnecessary effort. Dependability aggregates factors whether users feel in control of the interaction. Further, stimulation assesses how exciting and motivating it is to use the app, while novelty asks for the degree of innovation. The scales of the UEQ can be grouped into pragmatic quality (includes the dimensions perspicuity, efficiency, dependability) and hedonic quality (stimulation, novelty). Pragmatic quality aggregates task related quality aspects; hedonic quality the non-task related quality aspects.

Further, we compared the measured user experience of SERMO to results of other established products using a benchmark data set containing quite different products. The UEQ offers such a benchmark, which contains the data of 246 product evaluations with the UEQ (with a total of 9905 participants in all evaluations) [39].

We targeted to include at least five persons into the usability test. Previous studies from the human-computer interface literature have found that 80% of usability problems can be detected with only five research subjects [40]. Turner et al. even claim that the most serious usability problems can be revealed with only three subjects [41]. Participants were recruited amongst students and employees of two universities. Patients suffering from mental diseases were recruited from a clinic specialised in psychiatric disorders with in- and outpatient care. Psychologists of this clinic selected candidates to be asked to participate in the test to ensure that the people are mentally stable. Further, psychologists and psychotherapists were asked to participate in the testing. None of the participants who joined the test had been involved in the development of SERMO. Before the survey was conducted and participants were recruited, a clarification of responsibilities has been sent to the Cantonal Ethics Committee. The Ethics Committee confirmed that for the planned evaluation no approval is required.

5.2 Usability test results

The usability test took place between September 15 and October 24, 2019 and was performed by 21 persons (13 female, 8 male). 9 persons were currently under treatment in the clinic. None of them was currently keeping an emotion diary, but five confirmed to have problems in regulating their emotions. Furthermore, 4 psychologists and psychotherapists and 8 persons with different background (design, computer science, commercial, communication) from two universities joined the test. The age of the participants ranged between 22 and 67 (average 38.4). The average self-perceived technical competence was 7.3.

Figure 9 and 10 present the results of the judgements of all participants aggregated into the 6 categories. Findings show that the scores for attractiveness and two scales describing a pragmatic quality aspect (efficiency, perspicuity) are good, i.e., values are above 0.8 (see figure 9). In

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

particular, the participants confirm that the app is understandable, easy to learn, and clear (scale perspicuity) as well as friendly, attractive and pleasant (scale attractiveness, see figure 6). The scales describing hedonic quality (stimulation and novelty), i.e., fun to use, show neutral evaluations. When considering only the expert judgements, all mean values are good, i.e. a value above 0.8 is achieved for all scales (see figure 8).

Figure 7 shows the comparison of the judgements of all participants to the UEQ benchmark. For the categories perspicuity and efficiency a value above average was achieved. Attractiveness and novelty was judged as below average and dependability and stimulation is bad compared to the benchmark. According to the UEQ handbook, the label ″good″ means 10% of the results in the benchmark data set are better and 75% of the results are worse. The label ″Above average″ means 25% of the results in the benchmark are better than the result for the evaluated system, 50% of the results are worse. Interestingly, the four experts (psychologists or psychotherapists) judged the application more positively (see figure 8): The category attractiveness was good compared to the benchmark; perspicuity, efficiency and stimulation above average and novelty even excellent.

We observed a larger variance in the judgements of attractiveness, stimulation and novelty. The reliability of the UEQ scales attractiveness, stimulation and originality is good and excellent as the Cronbach-Alpha values indicate. Cronbach-Alpha is a measure of the internal consistency of a questionnaire dimension. The Cronbach-Alpha coefficients in our evaluation are 0.94 for attractiveness, 0.94 for stimulation and 0.85 for originality. Thus, these values indicate a good scale consistency. The Cronbach-Alpha values for efficiency (0.26), 0.53 for perspicuity, and dependability (0.27) are rather weak, i.e. the internal consistency is poor or even unacceptable. This can be due to problems with the interpretation of the items in these scales: Some UEQ items are difficulty to assess for SERMO (e.g. item ″slow/fast″ and ″not secure / secure″ cannot be judged by the users or interpreted differently).

During the interactions with the chatbot, the conversation sometimes stopped during the test due to unexpected user input. Conversation competencies of SERMO are still limited which was recognized by all participants. It became clear that even though the app is designed to collect the same information on a daily basis, the users desire a larger flexibility. Some users suggested reformulations of the chatbot responses to be more helpful and acceptable by the users.

6 Discussion

6.1 Lessons learnt from the usability test

Psychologists and psychotherapists confirmed that SERMO could be stimulating for patients. Obviously, a similar application is not yet in use in their daily treatment practice. They clearly see benefits of SERMO. We received the feedback from experts that the app is well suited for patients who have problems in expressing themselves in a face-to-face encounter. It could well bridge the gap between two therapeutic sessions (instead of calling the therapist, the patient could chat with SERMO). However, they suggested that a therapist should receive an alert when the system determines a certain risk for a patient from the conversations.

The results show that the tasks can be completed well with the app and it is easy to get used to the app (perspicuity and efficiency are good). By non-experts, the app is perceived as not very stimulating and motivating. This might be due to false expectations. Nowadays, people use voice user interfaces such as Siri, Alexa etc. The objective of those systems is to entertain and provide information. SERMO is not designed to entertain a user, but to collect information and improve certain skills of a user. The critical judgements regarding novelty, stimulation and attractiveness of the non-experts might also be due to their technical background: the participants were rather experienced.

The evaluation setting had several difficulties: The non-experts had a high technical competence, while the experts were rather inexperienced in technical issues (score average of 4). The non-experts were not informed on cognitive behaviour therapy and received only a brief introduction into the goals of SERMO. This might have impacted the expectations. We conclude that SERMO in its current stage still needs improvements with respect to design and variability of chatbot responses. The system has to be evaluated in a real world treatment setting where people are informed on the actual purpose of the app.

To study the user experience of SERMO, we decided to use the UEQ since a benchmark has been created that allows to better judge and compare the results with other systems. However, the benchmark does not reflect the peculiarities of systems or mobile applications in healthcare. There are other scales and questionnaires available such as the System Usability Scale. Chatbot usability is still a very incipient field as the study of Ren et al. shows [42]. Beyond usability and user experience, there are other aspects that have to be studied specifically for a health chatbot such as the specific task-oriented perspective and the clinical efficacy.

6.2 Comparison with existing mental health chatbots

SERMO is one of the first applications for supporting emotion regulation and processesing German natural language. There are several applications for mentally ill people on the market. Compared to the mental health chatbot Woebot for example [14], SERMO differs in the scope: While Woebot provides psychotherapy support and education, SERMO additionally aims to support practicing emotion regulation and allows self-monitoring of emotions and related events. Through automatic emotion recognition from free text user input by SERMO, the user can also be supported more specifically by appropriate information, tasks and exercises. The integrated feedback function on suggested exercises could help in future to study the effectiveness of different measures depending on individual situations and emotions. The system could also learn user preferences.

Another peculiarity of SERMO is that the system runs offline. The underlying technology is the OSCOVA framework. The OSCOVA NLP engine supports machine learning, i.e. OSCOVA trains itself to understand natural language which helps to improve the recognition rate for natural language user input. In contrast, Woebot is based on decision trees and is thus more restricted with respect to interpreting user input.

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

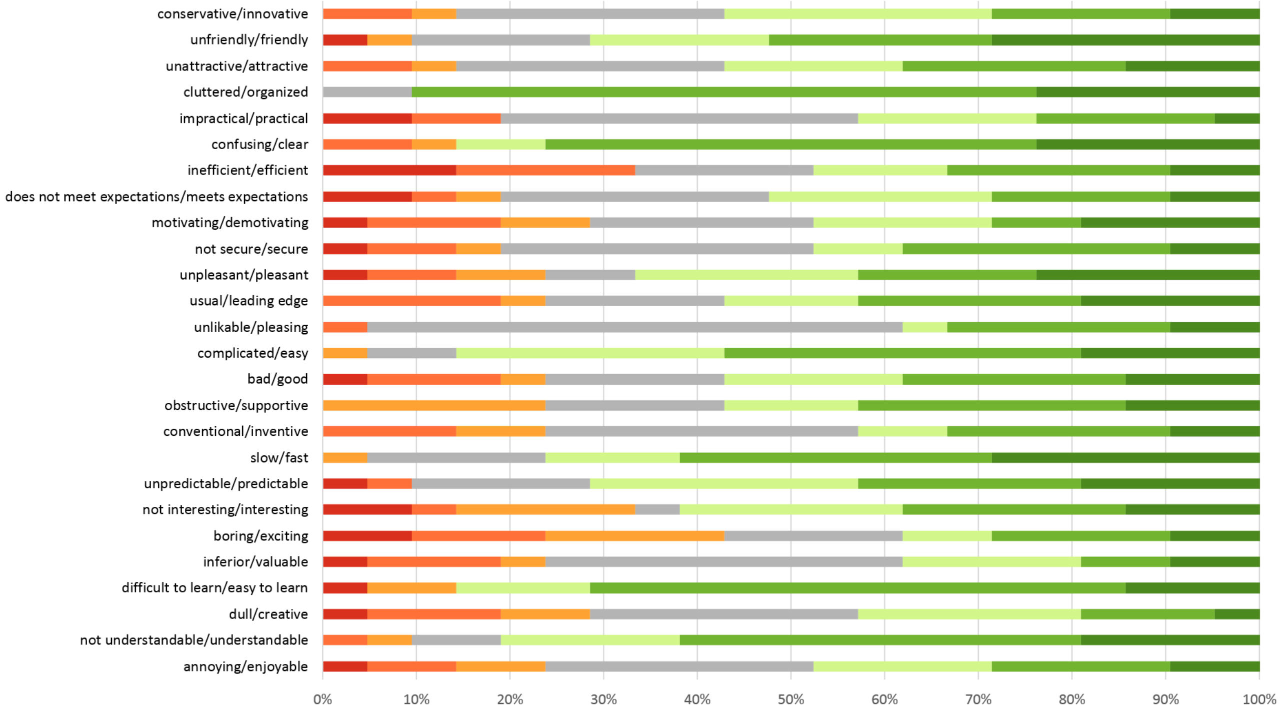


Fig. 6. UEQ Answers per item (n=21). -3 means fully agree with negative item, +3 means fully agree with positive item. Dark red (-3), light red (-2), orange (-1), grey (0), light green (+1), darker green (+2), dark green (+3)

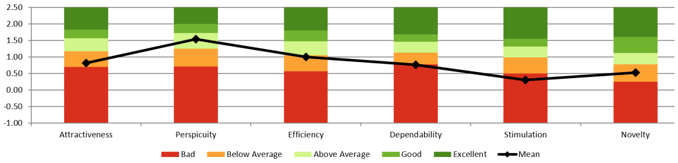


Fig. 7. Comparison of SERMO evaluation results with the UEQ benchmark (n=21)

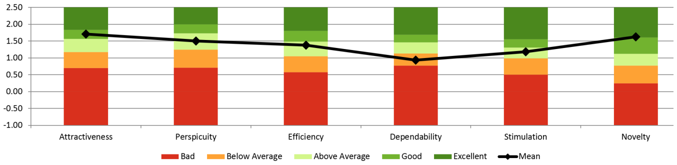


Fig. 8. SERMO evaluation results from psychologists and psychotherapists compared to the UEQ benchmark (n=4)

Confidence intervals (p=0.05) per scale						
Scale	Mean	Std. Dev.	N	Confidence	Confidence interval	
Attractiveness	0.810	1.407	21	0.602	0.208	1.411
Perspicuity	1.536	0.849	21	0.363	1.173	1.899
Efficiency	1.000	0.884	21	0.378	0.622	1.378
Dependability	0.762	0.900	21	0.385	0.377	1.147
Stimulation	0.298	1.656	21	0.708	-0.411	1.006
Novelty	0.524	1.313	21	0.562	-0.038	1.085

Fig. 9. Confidence intervals for the six categories (n=21)

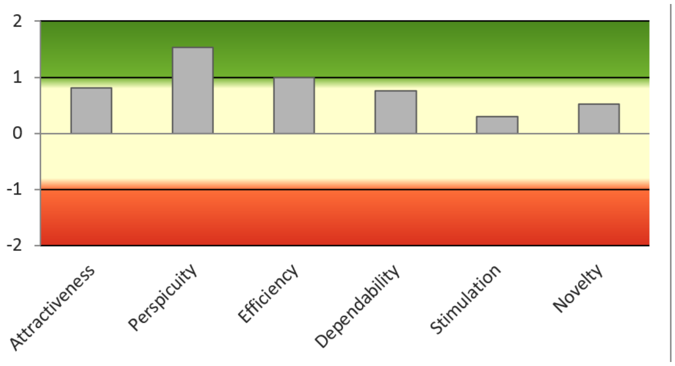


Fig. 10. Results of the UEQ questionnaire aggregated into six categories attractiveness, perspicuity, efficiency, dependability, stimulation and novelty (n=21)

In this paper, we provided a detailed description of SERMO, its knowledge base and its integrated technology. This will enable users and mental health service provider to inform themselves on the system and judge the evidence base and the reliability of chatbot answers. The system is not designed to generate responses on its own from training data. This is to ensure that the responses are reliable and the patient safety is not impacted. The system development was based on clinical evidences from CBT which are described in this paper. The conversation flow of SERMO was approved by psychologists. For many existing systems, it is unclear on which psychological principles they are based if any of them have been considered at all. We integrated exercises and knowledge retrieved from discussion with psychologists and from literature. Integrated exercises are currently only examples, i.e. the application has to be extended for a real-world use with additional exercises to increase the user experience and have a larger variability in suggested measures.

SERMO is not yet available in app stores as other mental

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

health chatbots. The reason is the current development phase: the system still needs improvements to be considered as the usability test showed.

6.3 Limitations of the emotion recognition

The implemented emotion recognition method still has potentials for improvement. First of all, the algorithm has to be extended to cover all relevant emotions. There are different theories on emotions and emotion types. We based our work on the theory of Berking and want to distinguish only seven emotions [17]. This was a result of our discussions with psychologists. Currently, the prerequisite for emotion recognition is that the user writes whole sentences in German, with punctuation marks, without errors. Smaller spelling errors can already be handled by the integrated fuzzy matching method. We have run a preliminary test on the emotion classifier with texts derived from a German depression forum. 50 statements were classified by SERMO. An accuracy of 81% could be achieved. Errors were partially due to the fact that the statements expressed emotions that SERMO is not yet able to determine. For six statements, no emotion was returned. Surprisingly, the algorithm well recognizes the emotions even though it is still simple. For example, the sentence “This does not make me angry, but sad” was classified correctly as “sadness”. However, the current lexicon-based approach for emotion classification still has limitations. A morphological analysis or at least a stemming algorithm could, among other things, help to improve the matching with the lexicon by reducing the terms of the user input to their lexical roots. In this way, the recognition rate could be improved. Furthermore, methods are necessary to determine implicit emotions. For example, when someone is writing “There is no sunshine inside of me” the person is most probably sad.

Existing work on emotion extraction from free text exploited support vector machines. Desmet and Hoste introduced an emotion classification algorithm and tested it on suicide note in English [43]. They distinguished 15 emotions. Their results show that the most salient features are trigram and lemma bags-of-words and subjectivity clues. Spelling correction had a slightly positive effect on classification performance. Shao et al. propose a lexicon-based emotion detection approach that combines basic grammars, tagged emotional words, and WordNet thesauruses [44]. Clearly, such existing work has to be considered for improving the emotion recognition in SERMO. Additional lexicons could be included such as LIWC for German [45] or GermaNet [46]. A future evaluation has to find out how reliably and correctly SERMO recognizes a user’s emotions. Another extension is to understand the entire context of emotion terms and in this way improve the emotion classification.

7 Conclusions and future work

In this paper, we introduced SERMO, a mobile application to support mentally ill people in regulating their emotions. The usability test results showed that the app is very well perceived by psychologist, but still needs improvements with respect to system stability, dealing with unexpected user input and variability of chatbots responses. In the current implementation, all data is stored on the mobile phone. In future, privacy issues have to be addressed. In case of a market-ready application, a usage condition, data protection declaration and declaration of consent covering all aspects of the General Data Protection Regulation has to be integrated. In the next phase of the project, we will work closely together with psychological experts and researchers to improve and expand the chat processes. Once the improvements of the app has been extended and the variability in chatbot answers has been increased, a pilot study will be conducted as randomized controlled trial with the target group to further study efficacy and usability of the app.

8 Acknowledgements

We would like to thank all persons participating in the usability test and the collaborating psychologists for their input and selection of patients to be involved in the testing.

References

[1] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, and D. Silove, “The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013,” *International journal of epidemiology*, vol. 43, no. 2, pp. 476–493, 2014.

[2] M. H. Foundation, “Fundamental facts about mental health,” *https://www.mentalhealth.org.uk/sites/default/files/fundamental-facts-15.pdf*, 2015.

[3] WHO, “Depression and other common mental disorders,” *Global Health Estimates*, 2017.

[4] H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, and T. Vos, “The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010,” *PloS one*, vol. 10, no. 2, p. e0116820, 2015.

[5] D. Schuler, A. Tuch, N. Buscher, and P. Camenzind, “Psychische gesundheit in der schweiz,” *Schweiz Gesundheitsobservatorium*, 2016.

[6] P. Cuijpers, M. Sijbrandij, S. Koole, G. Andersson, A. Beekman, and C. Reynolds, “The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons.” *World Psychiatry*, vol. 12, no. 2, pp. 137–48, 2013.

[7] C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla *et al.*, “Disability-adjusted life years (dalys) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010,” *The lancet*, vol. 380, no. 9859, pp. 2197–2223, 2012.

[8] B. D. Oladeji and O. Gureje, “Brain drain: a challenge to global mental health,” *BJPsych international*, vol. 13, no. 3, pp. 61–63, 2016.

[9] E. Anthes, “Mental health: theres an app for that,” *Nature News*, vol. 532, no. 7597, p. 20, 2016.

[10] R. D. Hester, “Lack of access to mental health services contributing to the high suicide rates among veterans,” *International journal of mental health systems*, vol. 11, no. 1, p. 47, 2017.

[11] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. S. Lau, and E. Coiera, “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 07 2018. [Online]. Available: https://doi.org/10.1093/jamia/ocy072

[12] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, and J. B. Torous, “Chatbots and conversational agents in mental health: A review of the psychiatric landscape,” *The Canadian Journal of Psychiatry*, 2019.

[13] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, “An overview of the features of chatbots in mental health: A scoping review,” *International Journal of Medical Informatics*, p. 103978, 2019.

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2974478, IEEE Transactions on Emerging Topics in Computing

[14] K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial,” *JMIR Ment Health*, vol. 4, no. 2, 2019.

[15] A. Beck, *Cognitive therapy and the emotional disorders*. International Universities Press, 1976.

[16] S. Barnow, “Emotionsregulation und psychopathologie,” *Psychol Rundsch.*, vol. 63, pp. 111–24, 2012.

[17] M. Berking, P. Wupperman, A. Reichardt, T. Pejic, A. Dippel, and H. Znoj, “Emotion-regulation skills as a treatment target in psychotherapy,” *Behav Res Ther.*, vol. 46, pp. 1230–7, 2008.

[18] A. Rathbone, L. Clarry, and J. Prescott, “Assessing the efficacy of mobile health apps using the basic principles of cognitive behavioral therapy: Systematic review,” *J Med Internet Res.*, vol. 19, 2017.

[19] A. Birney, R. Gunn, J. Russell, and D. Ary, “Moodhacker mobile web app with email for adults to self-manage mild-to-moderate depression: Randomized controlled trial,” *JMIR MHealth UHealth*, vol. 4, 2016.

[20] K. H. Ly, A.-M. Ly, and G. Andersson, “A fully automated conversational agent for promoting mental well-being: A pilot rct using mixed methods,” *Internet Interventions*, vol. 10, pp. 39 – 46, 2017.

[21] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, and I. Singh, “Can your phone be your therapist? young peoples ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support,” *Biomedical Informatics Insights*, vol. 11, 2019.

[22] B. Inkster, S. Sarda, and V. Subramanian, “An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study,” *JMIR Mhealth Uhealth*, vol. 6, no. 11, 2018.

[23] M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence,” in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE, 2017, pp. 858–862.

[24] K. M. Kudiri, G. K. Verma, and B. Gohel, “Relative amplitude based features for emotion detection from speech,” in *2010 International Conference on Signal and Image Processing*. IEEE, 2010, pp. 301–304.

[25] A. De and A. Saha, “A comparative study on different approaches of real time human emotion recognition based on facial expression detection,” in *2015 International Conference on Advances in Computer Engineering and Applications*. IEEE, 2015, pp. 483–487.

[26] A. Fernández-Caballero, A. Martínez-Rodrigo, J. M. Pastor, J. C. Castillo, E. Lozano-Monasor, M. T. López, R. Zangróniz, J. M. Latorre, and A. Fernández-Sotos, “Smart environment architecture for emotion detection and regulation,” *Journal of biomedical informatics*, vol. 64, pp. 55–73, 2016.

[27] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements,” *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[28] M. Skowron, “Affect listeners: Acquisition of affective states by means of conversational systems,” *Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 169–181, 2010.

[29] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional chatting machine: Emotional conversation generation with internal and external memory,” *Proceedings of the 2017 AAAI conference*, 2017.

[30] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013, pp. 1631–1642. [Online]. Available: https://www.aclweb.org/anthology/D13-1170

[31] K. Denecke and Y. Deng, “Sentiment analysis in medical settings,” *Artif. Intell. Med.*, vol. 64, no. 1, pp. 17–27, May 2015. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2015.03.006

[32] B. Wilken, *Methoden der Kognitiven Umstrukturierung. Ein Leitfaden fr die psychotherapeutische Praxis.* Verlag W. Kohlhammer, 1998.

[33] V. Vahia, “Diagnostic and statistical manual of mental disorders 5: A quick glance,” *Indian J Psychiatry*, vol. 55, pp. 220–23, 2013.

[34] C. Caldeira, Y. Chen, L. Chan, V. Pham, Y. Chen, and K. Zheng, “Mobile apps for mood tracking: an analysis of features and user reviews.” in *AMIA Annual Symposium proceedings. AMIA Symposium*, 2017, pp. 495–504.

[35] M. Linden, *Verhaltenstherapiemanual*. Springer Medizin Verlag, 2005.

[36] J. Brantley, “Mindfulness-based stress reduction,” *Acceptance and Mindfulness-Based Approaches to Anxiety: Conceptualization and Treatment*, pp. 131–145, 2005.

[37] R. Remus, U. Quasthoff, and G. Heyer, “SentiWS - a publicly available German-language resource for sentiment analysis,” in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10)*. Valletta, Malta: European Languages Resources Association (ELRA), May 2010.

[38] K. Chatzitheodorou, “Improving translation memory fuzzy matching by paraphrasing,” *Proceedings of the Work-shop Natural Language Processing for Translation Memories, Hissar, Bulgaria: Association for Computational Linguistics;*, pp. 24–30, 2015.

[39] M. Schrepp, A. Hinderks, and J. Thomaschewski, “Construction of a benchmark for the user experience questionnaire (ueq).” *IJIMAI*, vol. 4, no. 4, pp. 40–44, 2017.

[40] J. R. Lewis, “Sample sizes for usability tests: Mostly math, not magic,” *Interactions*, vol. 13, no. 6, pp. 29–33, Nov. 2006. [Online]. Available: http://doi.acm.org/10.1145/1167948.1167973

[41] C. Turner, J. Lewis, and J. Nielsen, “Determining usability test sample size,” *International Encyclopedia of Ergonomics and Human Factors*, pp. 3084–3088, 2006.

[42] R. Ren, J. W. Castro, S. T. Acuña, and J. de Lara, “Usability of chatbots: A systematic mapping study,” in *The 31st International Conference on Software Engineering and Knowledge Engineering, SEKE 2019, Hotel Tivoli, Lisbon, Portugal, July 10-12, 2019.*, A. Perkusich, Ed. KSI Research Inc. and Knowledge Systems Institute Graduate School, 2019, pp. 479–617. [Online]. Available: https://doi.org/10.18293/SEKE2019-029

[43] B. Desmet and V. Hoste, “Emotion detection in suicide notes,” *Expert Systems with Applications*, vol. 40, no. 16, pp. 6351 – 6358, 2013.

[44] Z. Shao, R. Chandramouli, K. Subbalakshmi, and C. T. Boyadjiev, “An analytical system for user emotion extraction, mental state modeling, and rating,” *Expert Systems with Applications*, vol. 124, pp. 82 – 96, 2019.

[45] T. Meier, R. L. Boyd, J. W. Pennebaker, M. R. Mehl, M. Martin, M. Wolf, and A. B. Horn, “liwc auf deutsch: The development, psychometrics, and introduction of de-liwc2015,” 2019.

[46] B. Hamp and H. Feldweg, “Germanet-a lexical-semantic net for german,” in *Automatic information extraction and building of lexical semantic resources for NLP applications*, 1997.

Kerstin Denecke is a professor of medical informatics at Bern University of Applied Sciences. Her research interests include medical language processing, information extraction, sentiment analysis, and text classification. Denecke received a Doctoral degree in computer science from the Technical University of Braunschweig. She is a member of the German Society of Medical Computer Science, Biometry and Epidemiology (GMDS); the German Journalists Association; and the IMIA Participatory Health and Social Media Working Group.

Sayan Vaaheesan studied medical informatics at the Bern University of Applied Sciences. In August 2019, he started as application developer at CISTEC AG.

Aaganya Arulnathan studied medical informatics at the Bern University of Applied Sciences. Since August 2019 she works as an IT consultant at ERNI Schweiz AG, a Swiss software engineering company.

2168-6750 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: UNIVERSITY OF BIRMINGHAM. Downloaded on May 10,2020 at 19:47:07 UTC from IEEE Xplore. Restrictions apply.