

Deep learning-based dimensional emotion recognition for conversational agent-based cognitive behavioral therapy

Julian Striegl^{Corresp., 1}, **Jordan Wenzel Richter**², **Leoni Grossmann**³, **Björn Bråstad**³, **Marie Gotthardt**⁴, **Christian Rück**³, **John Wallert**³, **Claudia Loitsch**¹

¹ Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI Dresden/Leipzig), Technische Universität Dresden, Dresden, Saxony, Germany

² Chair of Human-Computer Interaction, Technische Universität Dresden, Dresden, Saxony, Germany

³ Centre for Psychiatry Research, Department of Clinical Neuroscience, Huddinge & Stockholm Health Care Services, Region Stockholm, Karolinska Institute, Stockholm, Sweden

⁴ Kungliga Tekniska högskolan, Stockholm, Sweden

Corresponding Author: Julian Striegl

Email address: julian.striegl@tu-dresden.de

Internet-based cognitive behavioral therapy (iCBT) is a scalable, cost-effective, and low-threshold form of therapy. In recent years, the use of conversational agents such as chatbots and voice assistants for iCBT delivery has been investigated. Conversational agents can be utilized to recognize and track emotional states which can further be used for assessing progress during therapy, convey empathy, and eventually predicting long-term therapy outcome. Thus far, existing systems focus mainly on the use of categorical emotional approaches and thereby underestimate the complexity of emotional states. To allow for a more complete tracking of emotional states based on individual user utterances, we present a transformer model for dimensional text-based emotion recognition that has been fine-tuned with publicly available datasets. Results show that the fine-tuned model outperforms state-of-the-art for the specific emotion dimensions of valence (Pearson Correlation Coefficient $r=0.9$), arousal ($r=0.77$), and dominance ($r=0.64$) and shows good usability, acceptance, and empathic understanding in a conducted feasibility study with 20 participants.

Deep Learning-based Dimensional Emotion Recognition for Conversational Agent-based Cognitive Behavioral Therapy

Julian Striegl¹, Jordan Wenzel Richter², Leoni Grossmann³, Björn Bråstad³, Marie Gotthardt⁴, Christian Rück³, John Wallert³, and Claudia Loitsch¹

¹Center for Scalable Data Analytics and Artificial Intelligence, 01187 Dresden, Germany

²Chair of Human-Computer Interaction, TU Dresden, 01187 Dresden, Germany

³Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Huddinge, Sweden & Stockholm Health Care Services, Region Stockholm, 171 77 Stockholm, Sweden

⁴KTH Royal Institute of Technology, 100 44 Stockholm, Sweden

Corresponding author:

Julian Striegl¹

Email address: julian.striegl@tu-dresden.de

ABSTRACT

Internet-based cognitive behavioral therapy (iCBT) is a scalable, cost-effective, and low-threshold form of therapy. In recent years, the use of conversational agents such as chatbots and voice assistants for iCBT delivery has been investigated. Conversational agents can be utilized to recognize and track emotional states which can further be used for assessing progress during therapy, convey empathy, and eventually predicting long-term therapy outcome. Thus far, existing systems focus mainly on the use of categorical emotional approaches and thereby underestimate the complexity of emotional states. To allow for a more complete tracking of emotional states based on individual user utterances, we present a transformer model for dimensional text-based emotion recognition that has been fine-tuned with publicly available datasets. Results show that the fine-tuned model outperforms state-of-the-art for the specific emotion dimensions of valence (Pearson Correlation Coefficient $r=0.9$), arousal ($r=0.77$), and dominance ($r=0.64$) and shows good usability, acceptance, and empathic understanding in a conducted feasibility study with 20 participants.

1 INTRODUCTION

Digitized therapeutic approaches have been researched for decades. Empirical studies yielded promising results regarding acceptance and efficacy for treatments of depression and anxiety disorders (Etzelmueller et al., 2020). Furthermore, examinations comparing internet-based cognitive behavioral therapy (iCBT) with traditional face-to-face therapy indicated comparable therapeutic effects on disorder symptoms associated with depression and anxiety (Carlbring et al., 2018). In recent years, scientific attention has been directed towards incorporating conversational agents (CAs) such as chatbots and voice assistants into iCBT. While more clinical evidence is still needed, first studies indicate good effectiveness and acceptance by users (Abd-Alrazaq et al., 2019; Milne-Ives et al., 2020). CA-based CBT amalgamates the advantages of guided iCBT with unguided therapeutic strategies (cf. impact of guidance on internet-based mental health interventions by Baumeister et al. (2014)), affording the capacity for adaptive interventions and personalized treatment approaches (Mehta et al., 2021). Moreover, CA-based CBT offers substantial scalability and represents a cost-effective possibility for therapeutic intervention. Consequently, CA-based CBT can be deployed as an autonomous therapeutic regimen or a supplementary component alongside conventional therapeutic modalities.

The exploration of one's emotions is a central component of CBT, as is a strong foundation of trust between the therapist and the individual conveyed through empathy. Therefore, therapeutic chatbots

must be capable of recognizing and tracking the emotions of patients to methodically respond to and empathize with them effectively in therapeutic conversations. Research substantiated a correlation between interpersonal factors (empathic understanding, positive regard, and congruence) and positive therapeutic outcomes (Elliott et al., 2018). It was observed that interpersonal factors have an influence on the effectiveness of therapeutic interventions that often surpasses the effect of the treatment method itself (Lambert and Barley, 2001). Therefore, established guidelines for the communication between therapist and patient should be considered when developing CAs for CBT. Furthermore, next to expressing empathy, emotional states can be used to assess intrinsic goals, exhibited behaviors (Nelissen et al., 2007), and treatment effect (Hollon and Ponniah, 2010), and successful emotion regulation within a session can be used as a potential predictor of long-term therapy outcome (Mehta et al., 2021). The impact of empathic conversations and emotion recognition is therefore an important area for research and development of CA in the field of CBT.

While there is already promising work in the field of CA-based CBT systems with emotion recognition capabilities (Abd-Alrazaq et al., 2019), established systems use a categorical emotional recognition approach thereby limiting the complexity of tracked emotional states to predefined categories instead of the finer-grained inference of a dimensional output vector (Gabriels, 2019). Approaches for deep learning-based dimensional text-based emotion recognition have been proposed but thus far have not been used in the context of CA-based CBT and, moreover, could be improved upon by means of dataset merging and fine-tuning. Furthermore, while numerous studies have looked at the general acceptance and user satisfaction of CA-based CBT systems as a whole, the isolated acceptance of emotion recognition in this context and the perceived empathic understanding have thus far not been investigated. We address this limitation by introducing and evaluating a fine-tuned deep-learning approach for dimensional text-based emotion recognition for CA-based CBT. Model feasibility is tested in a user study, focusing on the isolated aspects, specifically, the acceptance of the performed emotion recognition, empathic understanding capabilities of the system, and the perceived appropriateness of connected system responses.

2 RELATED WORK

*Woebot*¹ is a mental health application that aims to provide content and techniques based on CBT to users via an integrated chatbot. Users can track their emotional states by selecting predefined emotional categories and by categorizing their moods through questions and answers. The tracking of emotional states, however, uses a categorical emotional model, thereby limiting the complexity of collected emotions (Gabriels, 2019). Fitzpatrick et al. (2017) investigated the acceptance and efficacy of the system with 70 participants in a randomized controlled study over a two-week period. Their results suggest a good acceptance and showed a significant decrease in symptoms of depression and anxiety in comparison to an e-book information control group. The acceptance and accuracy of the mood tracking functionality in the application have thus far not been investigated separately.

*Youper*² is a chatbot-based system providing exercises and content based on CBT to help people deal with emotional distress via emotion regulation approaches in an in-time intervention approach. Users can use a daily check-in functionality to track basic emotional states in a discrete categorical approach, combined with the possibility of recording an intensity level for the chosen emotional category. The acceptance and effect of the system on emotion regulation capabilities and symptoms of depression and anxiety were investigated in a longitudinal observational study with active customers of the platform by Mehta et al. (2021). Results indicate a good acceptance and a decrease in depression and anxiety symptoms after two weeks of usage. The acceptance was measured through a 5-star rating. Standardized questionnaires for determining acceptance, such as the Client Satisfaction Questionnaire for web-based health interventions (CSQi) (Boß et al., 2016) or the Net Promoter Score (NPS) (Baehre et al., 2022), were not taken into account. This impedes the assessment of the comparability of the study results with other systems.

The *Wysa*³ system combines chatbot-based therapeutic exercises with human mental health coaching. In the application, users can track their mood by choosing a predefined emotional category at the beginning of each session with the system. The application was used in multiple prospective cohort studies with

¹Woebot Inc., <https://woebothealth.com/>, accessed 27.07.2023

²Youper, <https://www.youper.ai/>, accessed 27.07.2023

³Wysa, <https://www.wysa.com/>, accessed 27.07.2023

participants with symptoms of anxiety and depression (Leo et al., 2022; Inkster et al., 2018). Results showed high patient engagement and improvements in anxiety and depression scores among participants of the high-usage group when compared to the low-app-usage group. However, a separate evaluation of mood tracking is also missing in this study.

*Cass AI*⁴ is a system designed to deliver adaptive conversations based on expressed emotions and mental health concerns of users (Joerin et al., 2019). Users can interact with the system either via free text input or by selecting predefined answers. According to the company, the system can detect patterns in phrasing, typing length of sentences, and number of grammatical errors to reveal dependencies to different emotional categories. The system was used by Fulmer et al. (2018) in a study to investigate its efficacy in reducing symptoms of depression and anxiety in college students. The authors conducted a single-blind randomized user study with an information control group (receiving an educational e-book on the topic). Results showed a significant reduction in symptoms of anxiety and depression in the experimental group. Participants from the experimental group furthermore showed higher levels of engagement and user satisfaction than those from the control. The accuracy of detected emotions and the level of conveyed empathy of the system have thus far not been investigated.

Compared to applying categorical emotion recognition approaches in the context of CA-based CBT, there is no comparable research on using dimensional text-based emotion recognition. However, there is some work on dimensional text-based approaches to emotion recognition in other contexts, which will be discussed subsequently.

Park et al. (2021) fine-tuned a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018a) to predict valence, arousal, and dominance (VAD) scores from input sentences. As the number of available VAD mapping datasets is small, the authors transformed data from categorical emotion mappings (SemEval⁵, ISEAR⁶, and GoEmotions (Demszky et al., 2020)) using the NRC emotional dictionary (Mohammad and Turney, 2010). Results outperformed the previous state-of-the-art in accuracy ratings for each VAD value. Their results demonstrate the feasibility of the mapping approach from categorical to dimensional emotional data and the potential of fine-tuned BERT models for emotion recognition. However, the authors did not pool data sets for training and did not investigate the acceptance and feasibility of the model in the context of CA-based CBT. Yang et al. (2023a) employed the VAD affect representation for emotion recognition in conversations using cluster-level contrastive learning. In a similar approach, they used disentangled variational autoencoders (Yang et al., 2023b) for VAD-based emotion recognition based on conversation histories. With both models, new state-of-the-art results could be achieved for two datasets. However, their approach is only applicable if a conversation history is already available. This cannot be a precondition for use in the area of CA-based CBT, because ethical considerations mean that context-free individual messages are preferable.

In summary, it can be concluded that established CA-based CBT systems offer mood tracking features, but so far the majority use a categorical emotion recognition approach, which limits the complexity of the recorded emotional states to predefined categories. In addition, no isolated evaluations of emotion recognition have been carried out in studies of the systems, which means that the accuracy and acceptance by users cannot be assessed individually. Furthermore, the state of the art reveals that approaches for deep learning-based dimensional text-based emotion recognition have previously been proposed but have not yet been used and researched in the application area of CA-based CBT. Existing approaches could be enhanced by pooling further data sets and fine-tuning them, as we will demonstrate in this paper.

3 DIMENSIONAL EMOTION RECOGNITION FOR CA-BASED CBT

We designed and developed a CA-based CBT system capable of performing text-based dimensional emotion recognition on individual user utterances. The system is powered by a dialogue management component that comprises both textual (chatbot (CB)) and vocal (voice assistant (VA)) means of communication to ensure accessibility for as many people as possible, including people with disabilities. The concept of the emotion-sensitive CA is illustrated in Figure 1. For the VA, user input is first transcribed by a speech recognition component and converted into text. The dialogue management system then infers user intent, recognizes the emotional state of the user, and returns the respective system response (see Figure 1a). While the structure of the CA-based CBT session itself is static (i.e., the system responses are not

⁴X2AI, <https://www.cass.ai/x2ai-home>, accessed 28.07.2023

⁵SemEval task E-c, <https://www.kaggle.com/datasets/context/semEval-2018-task-ec>, accessed 22.08.2023

⁶ISEAR dataset, <https://paperswithcode.com/dataset/isear>, accessed 22.08.2023

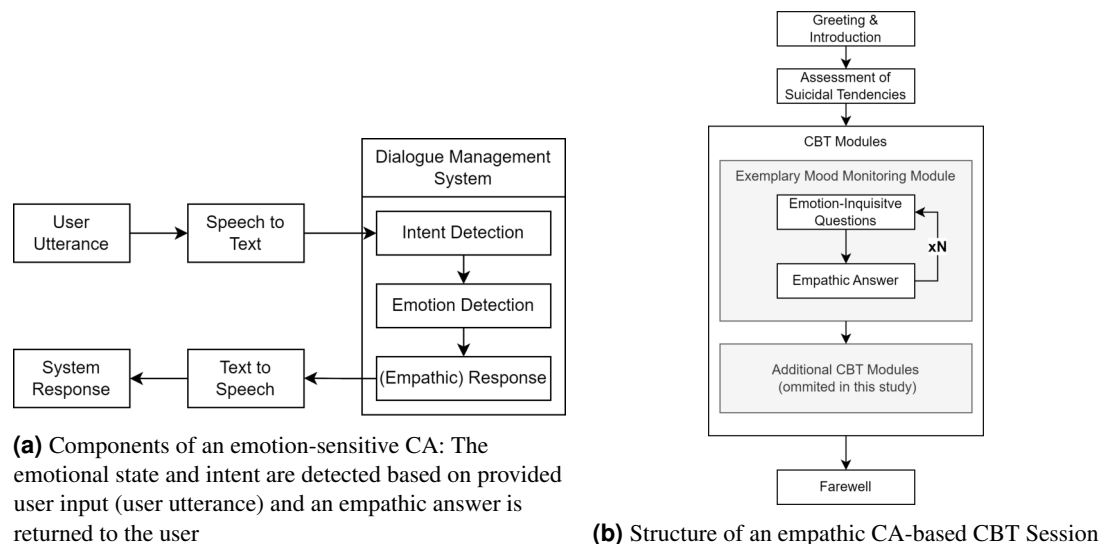


Figure 1. Concept of emotion-sensitive CA-based CBT

generated but are predefined in a closed-domain dialogue approach) and can consist of several CBT-based exercises, custom empathic responses are delivered at certain system states to simulate therapist empathy by inferring upon the emotion and selecting from a set of manually tailored empathic responses in a decision-tree approach. Finally, the textual response is synthesized, transformed back to vocal data, and returned to the user.

3.1 Empathic Dialogue Management

To demonstrate the benefits of empathic system responses in the context of CBT and to evaluate its acceptance, we propose a modular CBT session, which can be flexibly adapted and augmented (see Figure 1b). First, users are introduced to the system and familiarized with the input modalities, session contents, requirements, and duration. Subsequently, the user's suicidal tendencies are assessed, with a referral to human-operated suicidal hotlines if confirmed. Following these preliminaries, a sequence of emphatically enhanced CBT modules can be arranged tailored to the patient's needs. An exemplary mood monitoring module is proposed with a range of emotion-inquisitive questions to leverage the empathic capabilities of the CA as much as possible. The questions are designed to elicit open and complex input from the user regarding their feelings and experiences in the past, present, and future without limiting themselves to basic emotional categories.

3.2 Dimensional Emotion Recognition

As the focus of this research is on the emotion recognition component of the conceptualized system, two approaches for accomplishing dimensional emotion detection from text are proposed: a deep learning (DL) based and an auxiliary rule-based approach for evaluation purposes.

As shown in Figure 2, the DL-based approach utilizes a BERT architecture (Devlin et al., 2018b) with a final regression layer for computing a dimensional output. Specifically, a pre-trained ALBERT model (Lan et al., 2019) is used, which represents each word and sub-word unit as a vector in a higher dimensional space, taking into account the surrounding context (non-linear mapping), is fine-tuned on emotion-annotated data, and learns to infer a dimensional score on input sentences in a linear regression layer.

The auxiliary rule-based approach utilizes the approach by Badugu and Suhasini (2017). It first checks for negations before using the dimensional emotion dictionary by Kušen et al. (2017) to look up the emotion score associated with each word of the input and aggregate the results into a final emotional score.

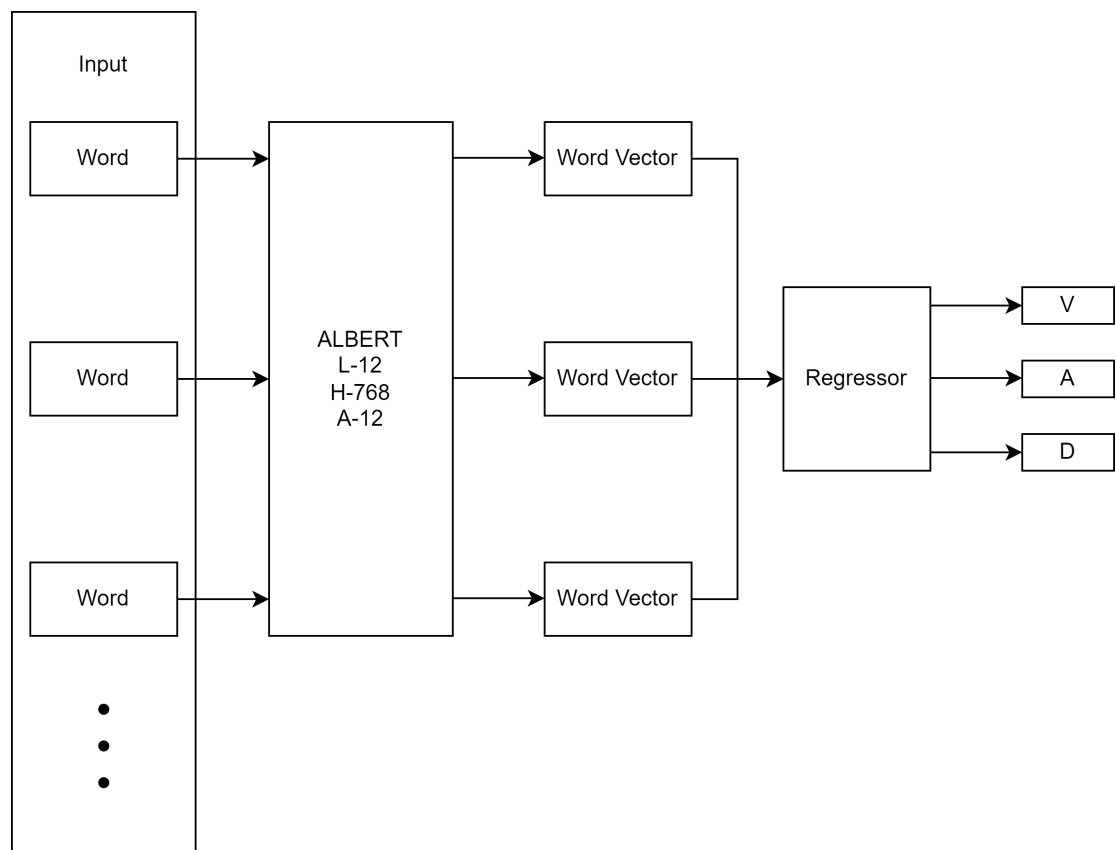


Figure 2. DL-based emotion recognition module.

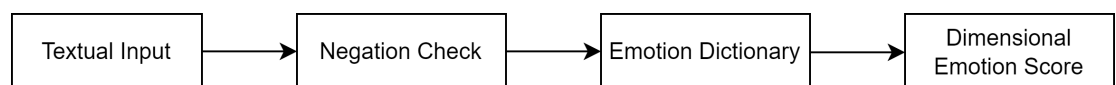


Figure 3. Rule-based emotion recognition module

3.3 Emotion Dataset Merging and Training

As most available emotion datasets are categorically labeled, a dataset transformation scheme is used (see Figure 4). Categorical labels in a given dataset are transformed into dimensional labels by taking the corresponding score in the NRC-VAD⁷, a dimensional emotion dictionary. The mean of the score is computed for multiple labels.

Four suitable datasets were identified and transformed if they were not inherently dimensionally annotated (see Table 1).

- The Emobank dataset (Buechel and Hahn, 2022) consists of over 10000 annotated English sentences, sourced from previously categorically annotated datasets (the manually annotated sub-corpus of the American National Corpus (Ide et al., 2008) and the SemEval-2007 Task 14 AffectiveText Corpus (Strapparava and Mihalcea, 2007)). Each sentence in the dataset was rated on its VAD value by five distinct human judges.
- The GoEmotions dataset (Demszky et al., 2020) consists of 58000 annotated comments of the social media platform reddit.com⁸, making it the largest emotion dataset annotated by humans currently. Each sentence in the dataset was labeled by three, and in cases of indecision, five human judges with one of 28 emotional categories.

⁷National Research Council Canada, <https://saifmohammad.com/WebPages/nrc-vad.html>, access date 13.08.2023

⁸Reddit, <https://www.reddit.com/>, accessed 23.11.2023

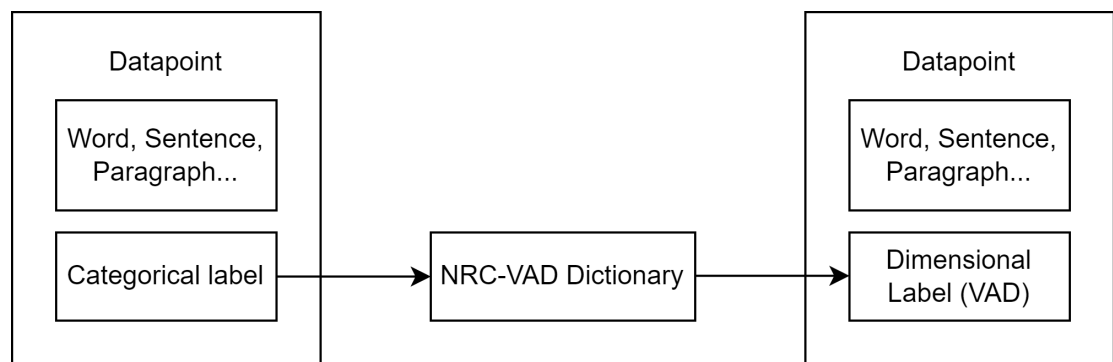


Figure 4. Transformation scheme from categorical to dimensional datasets

- The International Survey on Emotion Antecedents and Reactions (ISEAR)⁹ is an annotated dataset that comprises 7503 sentences, annotated for the emotional categories of joy, fear, anger, sadness, disgust, shame, and guilt by psychology students and non-psychology students.
- The CrowdFlower¹⁰ dataset comprised of over 7500 tweets, annotated by the emotional categories: empty, sadness, enthusiasm, neutral, worry, sadness, love, fun, hate, happiness, relief, boredom, surprise or anger. The dataset was exempt from training and reserved for testing to investigate the generalization ability of the trained model for different labeling heuristics. The CrowdFlower dataset was chosen for this as it has an uneven spread of emotional categories, making it more suitable for testing than training data.

Table 1. Comparison of datasets in regards to the amount of contained samples and the label type (native VAD labels, transformed categorical labels)

Dataset	Size	Labels
EmoBank	10000	native
GoEmotions	58000	transformed
ISEAR	7503	transformed
CrowdFlower	7500	transformed

The remaining datasets were merged to create one large dataset. As shown in Figure 5, combining EmoBank, GoEmotions, and ISEAR yielded a dimensional emotion dataset with 75503 samples and a more balanced VAD distribution than their individual constituents.

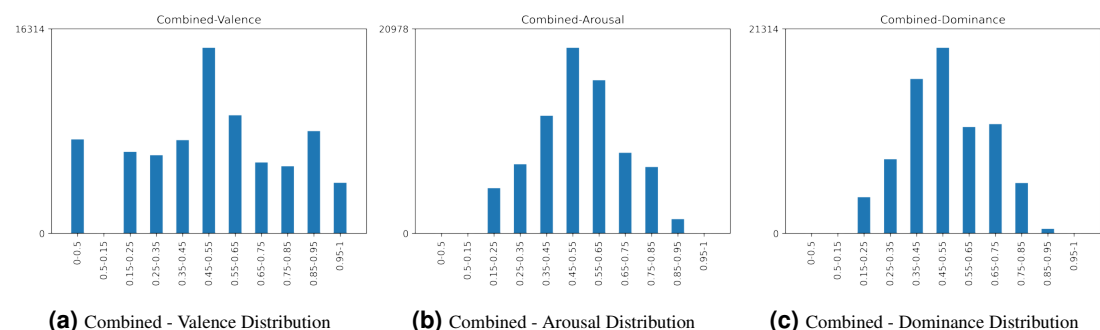


Figure 5. Combined data distribution for valence, arousal, and dominance

⁹ISEAR dataset, <https://paperswithcode.com/dataset/isear>, accessed 22.08.2023

¹⁰Sentiment Analysis in Text - Dataset by crowdflower, <https://data.world/crowdflower/sentiment-analysis-in-text>, accessed 07.08.2022

3.4 Model Development

To model the relationship between the textual input data and the dimensional emotion label, an adapted ALBERT Model has been trained on the transformed dataset. Hyper-parameter tuning and choosing of an appropriate ALBERT backbone were done using a train/validation/test split of 78.4%, 19.6%, and 2%, resulting in 59032, 14759, and 1506 samples respectively. The rather small test split still resulted in a sizeable sample size and was deemed sufficient for detecting overfitting of the validation set during hyperparameter tuning. The final model architecture comprised an input layer with variable input size, a preprocessing layer, a pre-trained ALBERT backbone¹¹ with 12 hidden layers/transformer blocks having the size of 768 units with GeLu activation and 12 attention heads, a dropout layer with dropout value 0.1 to prevent overfitting and a final dense layer with linear activation for predicting into the 3 VAD classes. This resulted in a total of 16 layers with 11,685,891 trainable parameters. The mean squared error was chosen as a reliable loss function for regression tasks, an adaptive learning rate of 3e-5, AdamW as the optimizer for penalizing large weights, and a batch size of 32.

3.5 Implementation

To use the trained model in the context of CA-based empathic CBT, a web application was implemented based on the previously developed concept (see Section 3.1). The user interacts with the CA via voice input or chat through a web-based user interface. The conversation with the CA is represented by speech bubbles to resemble a conversation with a human. A record button and an alternative text input field below the conversation can be utilized by the user to record answers to the CA. Open-source solutions were utilized for the web server, dialogue management, speech recognition, and speech synthesis, namely Flask¹², RASA¹³, Vosk¹⁴, and Rhasspy¹⁵.

4 EVALUATION

For evaluation, the chosen performance metrics of the fine-tuned model were compared to the alternative rule-based approach and to related work in the field. Additionally, a user study was conducted to assess the acceptance of the performed emotion recognition, empathic understanding capabilities of the system, and perceived appropriateness of empathic system responses in the context of CA-based CBT.

4.1 Technical Evaluation

In the following section, the quantitative results and used methodology of the technical evaluation will be described.

4.1.1 Methodology

Mean Squared Error (MSE) and Pearson Correlation Coefficient (r) were measured for both the DL-based and rule-based approach for comparative purposes. As the DL-based approach has been trained on most of the datasets, *Split* evaluates only the parts of the datasets that have not been used for training. Additionally, we compared with the results from Park et al. (2021), which used a similar dataset transformation scheme, but no dataset merging.

4.1.2 Results

As shown in Table 2 and Table 3, the DL model achieves smaller MSE and higher correlation throughout the datasets compared to the rule-based approach. It generally infers on the data more accurately, although the annotation procedure of the different datasets varies. When comparing the correlation coefficient to the state-of-the-art DL model for dimensional text-based emotion recognition by Park et al. (2021), the present model outperforms the one by Park et al. in all VAD dimensions (by $r=0.06$ for valence, $r=0.2$ for arousal, and $r=0.12$ for dominance, see Table 4).

4.2 User Study

A randomized A/B-testing experiment was conducted as an online between-subject feasibility study with healthy individuals.

¹¹ALBERT, <https://www.kaggle.com/models/tensorflow/albert/frameworks/tensorFlow2/variations/en-base/versions/2>, accessed 22.11.2023

¹²Flask, <https://flask.palletsprojects.com/en/3.0.x/>, accessed 23.10.2023

¹³RASA, <https://rasa.com/>, accessed 23.10.2023

¹⁴Vosk, <https://alphacephei.com/vosk/>, accessed 23.10.2023

¹⁵Rhasspy, <https://rhasspy.readthedocs.io/en/latest/>, accessed 23.10.2023

Table 2. Mean squared error for valence, arousal, and combined VAD-score for DL- and rule-based emotion recognition approach on different datasets

Mean Squared Error								
	DL-Group				Rule-Group			
	V	A	D	VAD	V	A	D	VAD
EmoBank	.0019	.0019	.0017	.0014	.0306	.0221	.0180	.0198
GoEmotion	.0042	.0041	.0029	.0029	.0532	.0358	.0270	.0319
ISEAR	.0011	.0013	.0011	.0008	.2125	.0974	.0734	.1104
Combined	.0036	.0036	.0026	.0025	.0661	.0401	.0304	.0381
Split	.0015	.0017	.0015	.0011	.2066	.1014	.0731	.1098
CrowdFlower	.0935	.0765	.0470	.0614	.1249	.0889	.0603	.0793

Table 3. The Correlation coefficient for valence, arousal, dominance, and combined VAD-score for deep learning- and rule-based emotion recognition approach on different datasets

Correlation Coefficient r								
	DL-Group				Rule-Group			
	V	A	D	VAD	V	A	D	VAD
EmoBank	.8952	.7674	.6436	.7687	.3627	.1863	.0749	.2080
GoEmotion	.9682	.9285	.9380	.9449	.4564	.1826	.2259	.2883
ISEAR	.9948	.9831	.9901	.9893	.3737	.0224	.2556	.2172
Combined	.9753	.9401	.9506	.9553	.4007	.1608	.1910	.2508
Split	.9925	.9785	.9861	.9857	.4118	-.0170	.2562	.2170
CrowdFlower	.5333	.2003	.4624	.3987	.2968	.1047	.1712	.1909

4.2.1 Methodology

Participants were semi-randomly assigned to two groups. Hence, group assignment was done randomly while ensuring equal group sizes for maintaining balance and comparability between groups. Both groups were led through an exemplary CBT session, with the only difference being the DL-based emotion detection in one and rule-based emotion detection in the other group. Participants needed to sign a privacy policy and consent form to comply with data protection provisions.

Demographic information, symptoms of depression using the short version of the Patient Health Questionnaire (PHQ2) (Löwe et al., 2005), and the affinity of technical interaction (ATI) (Franke et al., 2019) were measured as independent variables. Participants were asked to rate the perceived empathy, fluency, and relevance of system answers based on a 5-point Likert scale to assess the Empathic Understanding (EU) capabilities as proposed by Rashkin et al. (2018).

The usability of the system was assessed using the System Usability Scale (SUS) (Brooke, 1995) as it is one of the most popular and validated instruments for usability assessment (Bangor et al., 2008). The SUS investigates the perceived usability of a system with 10 questions based on a 5-point Likert scale, with the maximum score being 100 and a score above 68 being considered an above-average usability.

The Client Satisfaction Questionnaire adapted to Internet-based Interventions (CSQ-I) (Boß et al., 2016) was used to investigate the acceptance of the system as it has been developed and validated specifically for digital mental health interventions. Each item of the CSQ-I is scored between 1 and 5. For determining the overall acceptance rating of the respective subject, scores are summed up, therefore ranging from 8 (lowest) to 32 (highest), with 20 being the medium score.

4.2.2 Participants

20 participants (healthy individuals without a diagnosed mental health disorder) were recruited online and evenly split between the two groups (DL group and rule-based group). Age differences between the groups were not significant ($M_{DL}=34.7$, $SD_{DL}=12.45$, $M_{Rule}=27.7$, $SD_{Rule}=10.3$, $p = .21$). The differences between groups in terms of self-reported symptoms of depression (PHQ2) were not significant ($M_{DL}=2.3$, $SD_{DL}=1.85$, $M_{Rule}=1.6$, $SD_{Rule}=1.02$, $p = .33$). There were furthermore no significant differences in the measured technology affinity (ATI) ($M_{DL}=32.2$, $SD_{DL}=12.5$, $M_{Rule}=39.2$, $SD_{Rule}=8.68$, $p = .17$).

Half of the participants in the DL group and 40% in the rule-based group had prior experiences with VAs.

Table 4. Comparison of correlation coefficient results to results of Park et al. (2021) on EmoBank dataset

Correlation Coefficient r									
	Park et al.			DL-Group			Rule-Group		
	V	A	D	V	A	D	V	A	D
EmoBank	.84	.57	.52	.90	.77	.64	.36	.19	.07

4.2.3 Results

As shown in Table 5, questions concerning the participant's experience with the CA regarding EU showed no significant differences between groups in the combined score ($M_{DL}=10.5$, $SD_{DL}=2.25$, $M_{Rule}=10.9$, $SD_{Rule}=1.87$, $p=.67$). The Rule-group scored higher in the *Empathy/Sympathy* ($M_{DL}=2.9$, $M_{Rule}=3.3$) and *Fluency* ($M_{DL}=4.3$, $M_{Rule}=4.7$) questions, whereas the DL-group scored higher in the question regarding *Relevance* ($M_{DL}=3.3$, $M_{Rule}=2.9$).

In the SUS ($M_{DL}=72.5$, $SD_{DL}=12.9$, $M_{Rule}=77.8$, $SD_{Rule}=7.62$, $p = .31$) and the CSQ-I ($M_{DL}=72.5$, $SD_{DL}=12.9$, $M_{Rule}=77.8$, $SD_{Rule}=7.62$, $p = .31$) no significant differences could be established. Altogether, no significant differences could be found between the experimental groups with the EU, SUS, and CSQ-I questionnaires. Both approaches achieved good usability and acceptance scores and scored high in empathic understanding.

Table 5. User experience results for perceived empathic understanding (EU), system usability (SUS), and acceptance (CSQi) for the deep learning approach (DL Mean) and the rule-based control (Rule Mean) and standard deviation (SD)

Questionnaire	DL Mean	DL SD	Rule Mean	Rule SD
EU [3:15]	10.50	02.25	10.90	01.87
SUS [0:100]	72.50	12.89	77.75	07.61
CSQ-I [8:32]	14.00	07.95	16.40	06.97

5 DISCUSSION

To investigate the feasibility of the proposed model in the context of CA-based CBT, this study reports on the technical evaluation as well as on a conducted human feasibility test measuring the isolated acceptance and user satisfaction of the used emotion recognition, and perceived appropriateness of connected system responses. In the technical evaluation, the DL approach scored better than the rule-based approach in the metrics MSE and Pearson correlation coefficient. This finding was especially prominent when comparing the performance of data that were annotated under the same heuristics as it was trained on. When comparing performance on a foreign dataset, the CrowdFlower dataset, although the DL scored noticeably lower, it nevertheless outperformed the rule-based approach. In comparison to related work, the here presented fine-tuned DL model outperformed the results of Park et al. (2021) in all three inferred dimensions. The main difference between the approaches being that Park and colleagues did not use a combined dataset for training and evaluated their model with a test set of EmoBank only. The performance gap between datasets with known and unknown heuristics underlines the importance of training prospective DL models on multiple datasets to promote a generalized understanding of emotions.

In the user study, while no significant differences could be found between the two experimental groups regarding EU, SUS, or CSQ-I, the results indicate good usability and acceptance for both groups. Although the results of the study demonstrate both the technical feasibility and the usability and acceptance by users in the context of CBT, further implications for use in the field of iCBT need to be considered, which are discussed subsequently.

5.1 Implications of Deep Learning-based Dimensional Emotion Recognition for iCBT

Implementation of AI tools in the mental healthcare context presents both opportunities and challenges. When assessing previous research on AI in mental health care, it is clear that there are flaws in research methodology and quality, such as not reporting external validation, high risk of bias, and a lack of

transparency (Tornero-Costa et al., 2023). Ethical and legal issues come into play whenever automated machines are integrated in mental healthcare. When implementing AI in healthcare, Gerke S (2020) identified four primary ethical issues, which include, (1) informed consent to use, (2) safety and transparency, (3) algorithmic fairness and biases, and (4) data privacy, as well as five legal challenges: (1) safety and effectiveness, (2) liability, (3) data protection and privacy, (4) cybersecurity, and (5) intellectual property law. Automation of time-consuming tasks in iCBT could, through a positive lens, lead to improved cost-effectiveness, which is an important point in often over-encumbered and underfinanced psychiatry treatment and care contexts. The WHO has identified digital health, which includes AI, as a critical step in the development of making healthcare more efficient and accessible worldwide (Organization et al., 2023). Another potential benefit would be more consistency and thereby legal certainty on the patients' behalf. Even though contemporary iCBT consists of highly structured, RCT evaluated, treatment protocols that are executed by trained professionals, human therapists have limited memory and attention span, they will have "bad days", like one patient more than others, will have read one scientific paper but not another, possess differing skill and experience et cetera. All of the previously mentioned factors have the potential for varying the quality of the iCBT delivered. On the other side, variability in terms of human therapist performance may not matter that much in highly structured iCBT. Moreover, machines are not sentient beings and do not have a causal understanding of the patient and his or her world. This means that the machine is susceptible to producing errors in assessment and treatment that are in part fundamentally different from human errors. iCBT treatment is a complex task and automating aspects of it is still in its infancy. The consequences of this paradigm shift are predominantly understudied to date. Considering the aforementioned, below follows a set of hypothetical scenarios of varying degrees of automation in the iCBT context focused on emotion recognition and CA treatment with respect to clinical implementation.

Fully automated iCBT, including the prediction of emotional states coupled with a CA in charge of the iCBT with no human therapist involvement, would be both unwanted and unethical. For legal reasons, having a clinical professional involved and ultimately responsible for treatment is mandatory today and unlikely to change in the foreseeable future. Only hybrid solutions of man-machine co-involvement are therefore further discussed here. One such hybrid scenario would be the sole automation of emotion recognition. This scenario starts with initial machine recognition of emotional states derived from patients' responses as part of ongoing iCBT treatment. Estimated emotional states can then be fed to a human clinician as decision support. In theory, this could render an improved understanding of a patient's emotional state and also change of state across time during iCBT. This could ultimately improve treatment tailoring and effectiveness through the patient perceiving the therapist as more empathic, strengthening the therapeutic alliance. Furthermore, it would allow for modifications of ongoing therapy work modules to better suit the patient's emotional state. A potential risk with this approach would be the drift of the therapist's own emotional assessment influenced by the machine's estimated emotional state of the patient which may be wrong. An interesting but largely untested scenario would be extended automation of emotion recognition coupled with therapist-supported CA treatment. This would involve not only the potential benefit of emotion recognition discussed above but also cost-effective semi-automated treatment. One such implementation would be that the emotionally informed CA drafts empathically written therapy responses to the patient's messages and the human therapist then scrutinizes the responses and signs off on them with or without making prior changes. A major portion of iCBT costs come from therapists spending time drafting responses to patients in the treatment portal, unlocking a major potential for cost-saving strategies. An additional downside risk with this scenario would be that the human therapist – due to stress or other human factors – signs off on written responses of lower therapeutic quality. Proper training and structured follow-up of therapists are likely required in this scenario, which in turn may offset some of the cost-effectiveness of the approach. That stated since a major motivation for iCBT is cost-effectiveness, extending it with emotionally tailored CA seems in accordance with that overarching aim of iCBT.

5.2 Limitations

The present study includes several limitations in terms of the design choice and resource availability.

System Design First, the proposed system only imitated the beginning of a CBT session in order to investigate the isolated acceptance and appropriateness of recognized emotions and related empathic responses. Therefore, a follow-up study should investigate the acceptance and usability of a complete CA-based CBT session with an empathic agent using the presented approach for text-based emotion

recognition. Second, empathic responses were given by choosing from five pre-defined answers based on the detected VAD score. While the recommendations of Holtforth and Castonguay (2005) and Elliott et al. (2018) concerning empathic responses in the therapy context were considered, no further investigation regarding the appropriateness of pre-defined answers was undertaken. While the sample size of 20 is too small to draw concise conclusions, this is a potential explanation for the relatively equal user study results for both experimental groups for SUS, CSQi, and EU. Future work should expand the system e.g. by DL-based response generation to better leverage the advantages of dimensional emotion recognition in comparison to categorical approaches, making fine-grained tailored empathic answers possible.

Emotion Recognition Approach and Model Training Additionally, the presented approach focused on text-based emotion recognition of individual user utterances and did not take the history of the conversation into account. Therefore, based on the merged dataset a model for emotion recognition in conversations could be trained similar to the works of Yang and colleagues (Yang et al., 2023a,b). The model could be, furthermore, tested on additional datasets, as the main test set in this study (CrowdFlower) had a varying labeling quality, thereby possibly influencing the results of the technical evaluation of the model. Regarding the training of the model, a relatively small parameter amount and training time was chosen, due to resource limitations. As the performance of pre-trained large language models usually increases with training time and network parameters used for fine-tuning (Lan et al., 2019), those factors should be increased for future versions of the system. Furthermore, as the evaluation showed lower scores on test datasets with unknown heuristics, it can be hypothesized that training on more heuristically distinct datasets could have increased the generalization ability of the model. Unfortunately, there are limited appropriate emotion-annotated datasets available.

Significance and Comparability As we wanted to test the system's feasibility in a user study in an early stage of development, a small sample size of 20 participants was used thereby reducing the statistical significance and representativeness found results. As there have been thus far no other user studies investigating the isolated effect and acceptance of dimensional emotion recognition and empathic system responses in the context of CA-based CBT, there is a difficulty in comparing found results with related work. As discussed earlier, related work up to now merely investigated the acceptance and effect of CA-based CBT systems as a whole thereby making causal inference regarding the effects of empathic dialogue management difficult.

6 CONCLUSION

We presented a system for DL-based dimensional text-based emotion recognition for CA-based CBT. In comparison to a rule-based approach, the presented system showed considerably higher scores in a technical evaluation and surpassed the state-of-the-art for text-based emotion recognition on individual user utterances. A conducted user study investigating the acceptance, usability, and empathic understanding of the developed system showed no significant differences between DL- and rule-based emotion recognition and connected empathic responses, probably due to the use of pre-defined answers. Results for both user groups showed good scores for usability, acceptance, and empathic understanding. Therefore, the presented model should be used in follow-up studies in combination with a more elaborate empathic response generation, a complete CA-based CBT session, and a larger sample size. Furthermore, the presented model could be improved through additional training datasets and a longer training time. Additionally, a different large language model could be used as a basis for fine-tuning and a future iteration of the system could take the conversation history into account.

REFERENCES

- Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., and Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.
- Badugu, S. and Suhasini, M. (2017). Emotion detection on twitter data using knowledge base approach. *International Journal of Computer Applications*, 162(10).
- Baehre, S., O'Dwyer, M., O'Malley, L., and Lee, N. (2022). The use of net promoter score (nps) to predict sales growth: insights from an empirical investigation. *Journal of the Academy of Marketing Science*, pages 1–18.

- 417 Bangor, A., Kortum, P. T., and Miller, J. T. (2008). An empirical evaluation of the system usability scale.
418 *Intl. Journal of Human-Computer Interaction*, 24(6):574–594.
- 419 Baumeister, H., Reichler, L., Munzinger, M., and Lin, J. (2014). The impact of guidance on internet-based
420 mental health interventions—a systematic review. *Internet Interventions*, 1(4):205–215.
- 421 Boß, L., Lehr, D., Reis, D., Vis, C., Riper, H., Berking, M., Ebert, D. D., et al. (2016). Reliability and
422 validity of assessing user satisfaction with web-based health interventions. *Journal of medical Internet
423 research*, 18(8):e5952.
- 424 Boß, L., Lehr, D., Reis, D., Vis, C., Riper, H., Berking, M., and Ebert, D. D. (2016). Reliability and
425 Validity of Assessing User Satisfaction With Web-Based Health Interventions. *J Med Internet Res.*,
426 18(8):e5952.
- 427 Brooke, J. (1995). Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189.
- 428 Buechel, S. and Hahn, U. (2022). Emobank: Studying the impact of annotation perspective and represen-
429 tation format on dimensional emotion analysis. *arXiv preprint arXiv:2205.01996*.
- 430 Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., and Hedman-Lagerlöf, E. (2018). Internet-based vs.
431 face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic
432 review and meta-analysis. *Cognitive behaviour therapy*, 47(1):1–18.
- 433 Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions:
434 A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.
- 435 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018a). Bert: Pre-training of deep bidirectional
436 transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 437 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018b). Bert: Pre-training of deep bidirectional
438 transformers for language understanding.
- 439 Elliott, R., Bohart, A. C., Watson, J. C., and Murphy, D. (2018). Therapist empathy and client outcome:
440 An updated meta-analysis. *Psychotherapy*, 55(4):399.
- 441 Etzelmüller, A., Vis, C., Karyotaki, E., Baumeister, H., Titov, N., Berking, M., Cuijpers, P., Riper, H.,
442 and Ebert, D. D. (2020). Effects of internet-based cognitive behavioral therapy in routine care for adults
443 in treatment for depression and anxiety: systematic review and meta-analysis. *Journal of medical
444 Internet research*, 22(8):e18100.
- 445 Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young
446 adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot):
447 a randomized controlled trial. *JMIR mental health*, 4(2):e7785.
- 448 Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: development
449 and validation of the affinity for technology interaction (ati) scale. *International Journal of Human-
450 Computer Interaction*, 35(6):456–467.
- 451 Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., Rauws, M., et al. (2018). Using psychological artificial
452 intelligence (tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR
453 mental health*, 5(4):e9782.
- 454 Gabriels, K. (2019). Response to “uncertainty in emotion recognition”. *Journal of Information, Commu-
455 nication and Ethics in Society*.
- 456 Gerke S, Minssen T, C. G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare.
457 *Artificial Intelligence in Healthcare*, pages 295–336.
- 458 Hollon, S. D. and Ponniah, K. (2010). A review of empirically supported psychological therapies for
459 mood disorders in adults. *Depression and anxiety*, 27(10):891–932.
- 460 Holtforth, M. G. and Castonguay, L. G. (2005). Relationship and techniques in cognitive-behavioral
461 therapy—a motivational approach. *Psychotherapy: Theory, Research, Practice, Training*, 42(4):443.
- 462 Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: The manually
463 annotated sub-corpus of american english. In *6th International Conference on Language Resources
464 and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).
- 465 Inkster, B., Sarda, S., Subramanian, V., et al. (2018). An empathy-driven, conversational artificial
466 intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods
467 study. *JMIR mHealth and uHealth*, 6(11):e12106.
- 468 Joerin, A., Rauws, M., and Ackerman, M. L. (2019). Psychological artificial intelligence service, tess:
469 delivering on-demand support to patients and their caregivers: technical report. *Cureus*, 11(1).
- 470 Kušen, E., Cascavilla, G., Figl, K., Conti, M., and Strembeck, M. (2017). Identifying emotions in social
471 media: comparison of word-emotion lexicons. In *2017 5th International Conference on Future Internet*

- 472 of Things and Cloud Workshops (FiCloudW), pages 132–137. IEEE.
- 473 Lambert, M. J. and Barley, D. E. (2001). Research summary on the therapeutic relationship and psy-
474 chotherapy outcome. *Psychotherapy: Theory, research, practice, training*, 38(4):357.
- 475 Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for
476 self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.
- 477 Leo, A. J., Schuelke, M. J., Hunt, D. M., Metzler, J. P., Miller, J. P., Areán, P. A., Armbrrecht, M. A., and
478 Cheng, A. L. (2022). A digital mental health intervention in an orthopedic setting for patients with
479 symptoms of depression and/or anxiety: feasibility prospective cohort study. *JMIR Formative Research*,
480 6(2):e34889.
- 481 Löwe, B., Kroenke, K., and Gräfe, K. (2005). Detecting and monitoring depression with a two-item
482 questionnaire (phq-2). *Journal of psychosomatic research*, 58(2):163–171.
- 483 Mehta, A., Niles, A. N., Vargas, J. H., Marafon, T., Couto, D. D., and Gross, J. J. (2021). Acceptability
484 and effectiveness of artificial intelligence therapy for anxiety and depression (youper): Longitudinal
485 observational study. *Journal of medical Internet research*, 23(6):e26771.
- 486 Milne-Ives, M., de Cock, C., Lim, E., Shehadeh, M. H., de Pennington, N., Mole, G., Normando, E.,
487 and Meinert, E. (2020). The effectiveness of artificial intelligence conversational agents in health care:
488 systematic review. *Journal of medical Internet research*, 22(10):e20346.
- 489 Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical
490 turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational*
491 *approaches to analysis and generation of emotion in text*, pages 26–34.
- 492 Nelissen, R. M., Dijker, A. J., and de Vries, N. K. (2007). Emotions and goals: Assessing relations
493 between values and emotions. *Cognition and Emotion*, 21(4):902–911.
- 494 Organization, W. H. et al. (2023). Digital health in the european region: the ongoing journey to
495 commitment and transformation. In *Digital Health in the European Region: the ongoing journey to*
496 *commitment and transformation*.
- 497 Park, S., Kim, J., Ye, S., Jeon, J., Park, H. Y., and Oh, A. (2021). Dimensional Emotion Detection
498 from Categorical Emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*
499 *Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for
500 Computational Linguistics.
- 501 Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. (2018). Towards empathetic open-domain
502 conversation models: A new benchmark and dataset.
- 503 Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the*
504 *fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.
- 505 Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazzeri, L., Traver, V., Novillo-Ortiz,
506 D., et al. (2023). Methodological and quality flaws in the use of artificial intelligence in mental health
507 research: systematic review. *JMIR Mental Health*, 10(1):e42045.
- 508 Yang, K., Zhang, T., Alhuzali, H., and Ananiadou, S. (2023a). Cluster-level contrastive learning for
509 emotion recognition in conversations. *IEEE Transactions on Affective Computing*.
- 510 Yang, K., Zhang, T., and Ananiadou, S. (2023b). Disentangled variational autoencoder for emotion
511 recognition in conversations. *IEEE Transactions on Affective Computing*.