Message Recommendation Strategies for Tailoring Health Information to Promote Physical Activities

Longdan Hao[1] [0000-0002-1700-0037], Stefan Goetze[2] [0000-0003-1044-7343], and Mark Hawley[1] [0000-0002-2349-4491]

[1] Centre for Assistive Technology and Connected Healthcare (CATCH), The University of Sheffield, Sheffield, UK
{lhao4,mark.hawley}@sheffield.ac.uk
https://catch.sites.sheffield.ac.uk/

[2] Dept. of Computer Science, The University of Sheffield, Sheffield, UK
s.goetze@sheffield.ac.uk

Abstract. In many behaviour change interventions, computer-tailored health information has proven to be more effective than general health information. However, the majority of these studies have only achieved small effect sizes and the effectiveness of computer-tailored health communication (CTHC) remains inconsistent across different populations and behaviours. Since most CTHC studies measure a behaviour difference (e.g., steps per day) or biological difference (e.g., blood pressure), it is challenging to determine whether the intervention's success is due to the quality of message tailoring or other factors (e.g., user interface design). This paper presents a study that assesses the performance of various algorithms for tailoring health information. These algorithms include a rule-based approach, based on behaviour change theories and machine learning algorithms. Despite limited data, the evaluated algorithms significantly outperform random message selection, achieving a 1.7-fold increase in precision for predicting participants' preferred messages, and a 1.38-fold improvement in overall accuracy for anticipating participants' preferences.

Keywords: computer tailored health information · recommendation system · behaviour change.

1 Introduction

Physical activity (PA) is beneficial for improving and maintaining people's health [1]. Regular PA is an important factor to increase people's quality of life (QoL) [2]. However, insufficient PA level is prevalent globally, increasing the risk of death [1]. As a form of health education, the communication of health information can promote people's health behaviour, such as increasing their PA level [3]. With computer technology, selecting and delivering health information to a large population based on their own situations is more practical and efficient [4–6].

1.1 Computer Tailored Health Communication (CTHC)

Computer tailored health communication (CTHC) uses computer-based platforms for personal data collection and processing and to provide tailored information to its users [4–6]. CTHC is fostered by the development of two main fields: stage-based behaviour change theories[3] and computer technologies [4–6]. Another reason that highlights the critical need for further research in the field of CTHC is the suboptimal outcomes stemming from behaviour change interventions that rely on generic health information.

In CTHC systems, individuals' information can be collected via objective monitoring devices, self-reported data (e.g. by questionnaires), and third-party databases with individual's previous personal records (e.g., health apps, clinical records, web surfing data) [5, 6]. The collected data is then processed for message selection. Two systems approaches for message selection are commonly used in CTHC: rule-based systems based on preset rules (more details are explained in Section 2.2), and recommendation systems (the popular algorithms used in RS are explained in Section 2.3) based on historical data. The selected message is then delivered via a message delivery channel, e.g. by print material, telephone, email, web or phone applications [6].

1.2 Behaviour change theories applied in CTHC

Message selection rules in the majority of rule-based CTHC studies are supported by behaviour change theories [8]. Behaviour change theories are frameworks that aim to explain the underlying mechanisms and processes involved in human behaviour change and decision-making. These theories provide insights into the factors that shape an individual's behaviour, including cognitive, emotional, social, and environmental components. By identifying the key determinants of behaviour and the relationships between them, these theories help inform the development of effective strategies and interventions to promote positive behaviour change in various contexts, such as health promotion, education, and environmental conservation [9]. The most prominent behaviour change theories commonly applied in CTHC are transtheoretical model (TTM), Health Belief Model (HBM), Social cognitive theory (SCT) and theory of planned behaviour (TPB) [6, 10].

The *stage of change* (SoC) as defined in the TTM [11, 12], is the most prevalent behaviour change model[4] applied in the construction of personalised communication [6, 10]. The Stage of change (SoC) theory assumes that behaviour changes in a cyclical process [11, 12]. Each cycle contains *six stages of change*:

[3] Stage-based theories assume that psychological development and behaviour changes can be described by different continuous phases [7]. Further information is provided in Section 1.2

[4] Behaviour change models are more specific and practical tools that are derived from behaviour change theories. These models provide a structured approach to designing and implementing behaviour change interventions. They offer step-by-step processes and techniques to initiate and sustain behaviour change

1. *pre-contemplation* (no intention to change)
2. *contemplation* (being considerate to change in the foreseeable future)
3. *preparation* (starting to make plans for changes in the near future as well as measurable preparations)
4. *action* (changes happened within the past six months)
5. *maintenance* (keeping and preserving changes for more than six months)
6. *termination* (having no intentions to return to the unhealthy behaviour nor relapse)

The *SCT* assumes that behaviour change is influenced by an individual's characteristics, environment, and by the behaviour itself. Adoption of behaviour change is seen as an interactive and reciprocal process. *Self-efficacy* is considered the most critical influence of behaviour change in SCT. In addition, the likelihood of an individual engaging in a behaviour change is also influenced by outcome expectancy [12]; The *health belief model (HBM)* assumes that the perceived threats of the behavioural consequences influence the adoption or cessation of behaviour. Perceived threats consist of two components: susceptibility, such as e.g. the risk of an individual getting a disease, and severity, e.g. how severe a disease could develop [12];

In the *TPB*, positive attitudes, perceived normative pressure, and perceived behaviour control are positively associated with greater behaviour changes. Perceived behaviour control presents the extent of an individual's beliefs in the user's ability to control and manage a behaviour change [12, 13]

In an effort to make behaviour studies more replicable and implementable, Michie et al. [14] consolidated the *active ingredients* used in behaviour change interventions into a unified reporting language. These behaviour change techniques (BCTs) are summarised in a respective BCT taxonomy [14] and are specific tactics or methods designed to facilitate individual behaviour modification. A total of 93 BCTs were identified and defined in [14]. Repeating all 93 BCTs

Table 1. Main BCT categories from [14] and number of respective sub-categories.

BCT main category (# sub.-cats.)	BCT main category (# sub.-cats.)	BCT main category (# sub.-cats.)
1. Goals and planning (9)	7. Associations (8)	13. Identity (5)
2. Feedback and monitoring (7)	8. Repetition and substitution (7)	14. Scheduled consequences (10)
3. Social support (3)	9. Comparison of outcomes (3)	15. Self-belief (4)
4. Shaping knowledge (4)	10. Reward and threat (11)	16. Covert learning (3)
5. Natural consequences (6)	11. Regulation (4)	
6. Comparison of behaviour (3)	12. Antecedents (6)	

identified in [14] is beyond the scope of this work but since the proposed algorithms described in Section 2 will be based on these BCTs, Table 1 summarises their main categories and how many sub-categories each main category has. The interested reader is referred to [14][5] for further details.

[5] A list of the 93 BCTs [14] is also available only at `https://digitalwellbeing.org/wp-content/uploads/2016/11/BCTTv1_PDF_version.pdf`, Last access 23/06/2023.

1.3 Effectiveness and challenges of CTHC

CTHC has been applied to promote different health behaviours to different populations, such as promoting disease screening behaviour [15], PA promotion [16], and nutrition promotion [17, 18].

A meta-analysis [10] assesses the efficacy of computer-tailored *printed* materials on health behaviour changes and finds that CTHC is effective (but with an effective size less than *small*). Furthermore, the study finds that a larger effect size is seen in (i) studies using a combination of different behaviour change theories, (ii) studies targeting multiple behaviours, and (iii) tailoring on demographic features [10].

Another meta-analysis [19] assesses the efficacy of computer-tailored *web-delivered* health messages. Overall, tailored messages on behaviour changes show a greater effect compared to control conditions (with a small effect size). Seven health behaviours were targeted in the CTHC interventions: PA, nutrition/diet, smoking, drinking, medication adherence, stress management, and faecal soiling. Within the included studies, tailored interventions were more effective for the healthy population than for people with conditions [19].

The meta-analysis [20] assesses the efficacy of CTHC on promoting PA to people with or at risk of long-term conditions. Overall, tailored messages on PA promotion show a greater effect compared to no health information and general health information. The overall effect size was small to medium. The assessed conditions include cardiovascular diseases, cancer, diabetes, COPD, overweight and obesity.

Despite previous studies demonstrating the effectiveness of CTHC, the overall effect size has been found to be small [10, 19, 20]. The tailoring rule is a key component in CTHC that influences the effectiveness of behaviour change interventions. As a result, this study primarily focuses on evaluating the accuracy of tailoring rules using different algorithms.

This study has two primary aims. The first is to assess the feasibility of using a hybrid recommendation system, combining a predefined rule-based approach and machine learning (ML)-based algorithms, to deliver tailored health information related to PA via a web application. To evaluate the performance of the different tailoring algorithms, we will compare confusion metrics (cf. Table 2) between the system's predicted rating scores[6] Y_p and the actual rating scores Y. The second aim is to enhance the system's usability to be addressed in future development stages.

2 System Description

Most previous CTHC systems employ rule-based approaches, utilizing a set of predefined rules grounded in behaviour change theories [21]. These rules involve

[6] rating scores are collected through a rating system to record participants' attitudes on messages. For details, see Section 3.2.

choosing suitable tailoring variables (e.g., individuals' stage of change), selecting relevant behaviour change theories (e.g., the stage-of-change model), and establishing conditional "if-else" selection processes [21, 22]. In CTHC, both individuals and messages are categorised into subgroups (also known as segmentation) based on the selected independent variables and predetermined rules [21]. If the message's subgroup aligns with the individual's subgroup, the message will be recommended [21]. Recently, research has shifted away from rule-based systems, with an increasing focus on exploring the application of recommendation systems (RSs) utilizing ML techniques to deliver tailored health communication [23].

The proposed CTHC system comprises, therefore, a rule-based algorithm (see Section 2.2) and machine learning algorithms (see Section 2.3). Their performance will be compared in Section 4.

2.1 System framework

The CTHC system proposed in this study consists of three main modules: message delivery, message selection, and database. The *messages delivery module* is responsible for identifying users, presenting tailored messages to the user, and saving rating records from the user to the database. The *message selection module* employs two filters for message selection. The first filter consists of a rule-based algorithm informed by the TTM, the health action process approach (HAPA)[7] and SCT. The second filter comprises the RS algorithm as described in Section 2.3. This filter is trained using the dataset collected from participants, which is stored in the *database module* which stores all relevant data on users, messages and ratings.

2.2 Rule-based algorithm

The proposed rule-based algorithm is based on three aspects described in the following to produce a matching score $0 \leq M_{u,m} \leq 1$, evaluating the relationship between each user u and each message m. The matching score $M_{u,m}$ depends on:

(1) The difference between a user's stage of change regarding PA, defined as an integer value $S_{\mathrm{u}} \in \{1, ..., 6\}$, and the stage of change best suited for a given message, defined accordingly as $S_m \in \{1, ..., 6\}$, cf. (1)
(2) The required BCTs for the user at their stage of change, and the BCTs associated with a message, cf. (2.2)
(3) Whether the message contains information about the user's general preferences and barriers to perform PA, cf. (7)

[7] HAPA posits that health behaviour change occurs in two distinct phases: the motivation phase and the volition phase [24, 25].

(1) User and Message Stage of change Difference: The SoC difference between user u and message m is denoted by $S_{u,m}$ and will be explained in the following, see. (1). The user's stage of change S_u is obtained by a baseline assessment detailed in Sections 3.2 and 4.1. Messages can be categorized to fit users with different stages of change based on the message's topic and emphasis on BCT. For example:

- Messages aimed at raising consciousness for behaviour change are more effective for people in SoC 1 [11, 12].
- Messages focused on problem solving (addressing barriers to physical activity) are more effective for people in SoC 2 (general barriers to performing PA) and SoC 3 (specific barriers to performing PA), so either a SoC of 2 or 3 is chosen for the SoC S_m best fitting for message m [12].
- Messages that convince people to create action plans are more effective for individuals in SoC 3 [24, 25].
- Messages that encourage self-monitoring and goal-setting are more effective for people in SoC 4 [11, 24, 25].

Hence, each message is assigned a SoC label S_m by manual annotation. A (weighted) absolute difference between the SoC $S_u \in \{1, ..., 6\}$ of user u and the SoC $S_m \in \{1, ..., 6\}$ which is best suited for message m is calculated by

$$S_{u,m} = \beta_s \cdot |S_u - S_m|. \tag{1}$$

When user u has a higher SoC S_u compared to the message's labelled SoC S_m, the impact of the difference is less significant in (1) than when user u has a lower SoC compared to the message's labelled SoC. To represent this difference, the weighting factor

$$\beta_\mathrm{s} = \begin{cases} 0.05 & \text{for } S_u \geq S_m \\ 0.1 & \text{for } S_u < S_m \end{cases} \tag{2}$$

is defined. Values 0.05 and 0.1 in (2) have been determined empirically since the value of the impact of the difference is unknown from previous studies.

(2) Matching Behaviour Change Techniques (BCTs) between Users and messages: BCTs are associated to each user depending on (i) the user's *SoCs*, (ii) their current *PA status* and a *psychological assessment* (self-efficacy on PA). The rules for labelling are supported by the TTM, the HAPA, and SCT.

The BCT taxonomy [14] (cf. Table 1) contains 16 main BCT categories ($1 \leq i \leq 16$) and a varying number $1 \leq j \leq J_i$ of techniques within each of these main categories, with J_i representing the number of BCTs within a main category i (cf. numbers in Table 1). Hence, a BCT vector

$$\mathbf{b} = \Bigg[\underbrace{B_{1,0}}_{\text{cat. 1}}, \underbrace{B_{1,1}, ..., B_{1,J_1}}_{\text{BCTs in cat. 1}}, \underbrace{B_{2,0}}_{\text{cat. 2}}, \underbrace{B_{2,1}, ..., B_{2,J_2}}_{\text{BCTs in cat. 2}}, ..., \underbrace{B_{16,0}}_{\text{cat. 16}}, \underbrace{B_{16,1}, ..., B_{16,J_{16}}}_{\text{BCTs in cat. 16}} \Bigg]^T \tag{3}$$

of length $93 + 16 = 109$ is defined, with all $B_{i,j} \in \mathbb{B}$, i.e. all entries of $\mathbf{b}$ are Boolean expressions $\mathbb{B} \in \{0, 1\}$. Operator $[]^T$ denoted the vector transpose.

In addition to identifying the required BCTs for each user, HAPA also suggested that some BCTs are more effective in particular SoCs than in others. For example, promoting perceived self-efficacy (BCT category 13, *Identity*, cf. Table 1) is considered more important in SoC $S_u = 3$ and $S_u = 4$ than in SoC $S_u = 1$ and $S_u = 2$. Risk perception is suggested to be more important in SoC $S_u = 1$ than other SoC statuses. Therefore, the vector $\mathbf{b}$ of Booleans is multiplied by a weighting vector $\boldsymbol{\beta}_{\text{BCT},s}$ for each user for all SoCs S_u between 1 and 4. Since the quantitative effects of BCTs have not been reported in previous studies, three different weighting values for $\boldsymbol{\beta}_{\text{BCT},s}$ are defined empirically, i.e. 0.05 (important BCTs), 0.02 (moderately important BCTs) or 0 (less important BCTs). Based on this, each user is assigned a user vector that represents the BCTs and their respective importance for that individual:

$$\mathbf{u}_s = \boldsymbol{\beta}_{\text{BCT},s} \circ \mathbf{b} \tag{4}$$

The symbol $\circ$ denotes the Hadamard product, i.e. the element-by-element multiplication.

Similar to vector $\mathbf{b}$, a vector $\mathbf{m}$ of length 109 containing Boolean values $M_{i,j} \in \{0, 1\}$ is defined, By manual annotation of the messages, $M_{i,j}$ is set to 1 if the message content includes the corresponding BCT and to 0 otherwise. For example, if a message contains information for persuading individuals to make plans for walking $10,000$ steps per day, we set $M_{1,4} = 1$ (action planning). If this message also provides reasons for why 10,000 steps per day are needed, we set $M_{5,1} = 1$, since the reason is related to information about health consequences. Similarly, $M_{5,3} = 1$ if the reason is related to making friends, i.e. if the message provides information about social and environmental consequences. $M_{5,6} = 1$ if the message discusses mental health, i.e. provides information about emotional consequences. This annotation leads to a vector

$$\mathbf{m}_m = [M_{1,0}, ..., M_{1,J_1}, M_{2,0}, ..., M_{2,J_2}, ..., M_{16,0}, ..., M_{16,J_{16}}]^T \tag{5}$$

matching the size of 109×1 of $\mathbf{b}$ in (3) and (4). The user BCT labels from (4) are then matched with the BCT features of each message as defined in (5) by calculating the product of the two vectors. This results in a value

$$B_{u,m} = \mathbf{u}_s^T \cdot \mathbf{m}_m, \tag{6}$$

representing the BCT matching level of SoC between each user u and each message m.

(3) Keyword Matching: In addition to matching BCTs and SoC, the rule-based algorithm also takes into account the user's general preferences and barriers to performing PA. The system generates a keyword list for each user's preferences and barriers (from the initial user assessment, see Sections 3.2 and

4.1 for details), as well as a keyword list for each message's content (and subtitle, if the message contains a video). The message's keyword list is extracted using the rapid automatic keyword extraction algorithm (RAKE) algorithm[8]. Keywords from the two lists are combined into pairs. For each pair, a match between the user and the message is considered, if the *fuzzy string match score*[9], being a value between 0 to 100%, is greater than 50%. The number of matches is then weighted by the ratio score $\beta_{\mathrm{r}_i,u,m}$, and then summed up.

$$K_{u,m} = \begin{cases} \sum_{i=0}^{n} \lfloor \beta_{\mathrm{r}_i,u,m} \rfloor & \text{if } \beta_{\mathrm{r}_i,u,m} \geq 50\% \\ 0 & \text{if } \beta_{\mathrm{r}_i,u,m} < 50\% \end{cases} \tag{7}$$

In (7), n denotes the number of $\beta_{\mathrm{r}_i,u,m} \geq 50\%$.

Message-User Matching Score: The final score for the general *degree of matching* between a user u and a message m is then based on SoC matching in (1), BCT matching in (6), and keyword-matching in (7):

$$M_{u,m} = \frac{1}{1+e^{(S_{u,m}-B_{u,m}-K_{u,m})}} \tag{8}$$

The rationale behind the empirically determined sigmoid relation in (8) is that when there is a large difference, or number of matched BCTs and keywords between user u and message m, the change in the included features (SoC, BCTs, keyword matching) should have a smaller impact on $M_{u,m}$.

2.3 Machine Learning Algorithms for Message Selection

Since the rule-based approach described before relies on several empirically determined components, three different, relatively simple machine learning (ML) methods [26] are tested and applied for user-message matching in the following to predict users ratings, i.e. Naïve Bayes (NB), support vector machine (SVM), and k-nearest neighbors (KNN) classifiers.

Data Pre-Processing: Due to the limited number of rating records, the 5-star rating scale ($1 \leq y \leq 5$) is converted to Boolean values *like* (i.e. 1, if $4 \leq y \leq 5$) and *dislike* (i.e. 0, if $1 \leq y \leq 3$). After conversion, y is the Boolean value of the rating $y \in \{0, 1\}$ (cf. Figs. 3 and 4 later in this document).

To reduce the variance of input features, individuals' PA levels are converted to Boolean values, with 1 representing the individual reaching the recommended PA level and 0 indicating otherwise. The individuals' self-efficacy (ranging from 5 to 40), delta stage of change, and the number of keyword matches generated by "TheFuzz" algorithm (see Section 2.2) were normalised using min-max normalisation (cf. (9)).

[8] Rapid automatic keyword extraction algorithm Python library: `https://pypi.org/project/rake-nltk/`, last access 23/06/2023

[9] The fuzzy string match score is calculated using *TheFuzz* algorithm, `https://pypi.org/project/fuzzywuzzy/`, last access 23/06/2023

Rating records were obtained from a trial run using the previously described rule-based approach. Those data were divided into a training dataset (70%) and a test dataset (30%).

All variants of input feature vectors $\mathbf{x}$ introduced in the following are normalised by min-max normalisation before being used as input for the classifiers, i.e.

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - x_{\min}}{x_{\max} - x_{\min}}, \tag{9}$$

with $x_{\min}$ and $x_{\max}$ denoting the minimum and maximum values of input feature vector $\mathbf{x}$, respectively.

Cosine Similarity of Text Embeddings: The cosine-similarity

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{a}\|} \tag{10}$$

is widely used to evaluate pairwise distances of vectorised items (e.g., message representations) in the vector space model (VSM)[10] of documents. The cosine-similarity will be used in the following by the classifiers to compare two vectorised texts: $\mathbf{a}$ denotes the vectorised text of each user's preference in general and barriers to performing PA, and $\mathbf{b}$ is the respective vectorised message. In VSMs, information vectorisation is an important step before the information classification. Typical methods in the field of natural language processing (NLP) are bag-of-words with term frequency–inverse document frequency (Tf-Idf) representations. However, in recent years, approaches based on word embeddings, such as Word2Vec [27] or bidirectional encoder representations from transformers (BERT) [28], have been shown to be successful for various NLP-related tasks. In this work, BERT representations are therefore used to vectorise text from both users and messages, i.e. for vectors $\mathbf{a}$ and $\mathbf{b}$ in (10).

Input Feature Vectors: Different combinations of features are defined in the following as feature input vectors for the classifiers under test. First, a vector of dimension 3×1, composed of the user's stage of change S_u, the user's PA level P_u and the user's self-efficacy level on performing PA E_u is defined, all values are determined during baseline assessment as described in Sections 3.2 and 4.1.

$$\mathbf{x}_u = [S_u, P_u, E_u]^T \tag{11}$$

A second input feature vector of size 4×1, capturing the user-message relationships is defined, comprising the number of BCT matches between the user and the message[11], $\mathbf{b}^T \cdot \mathbf{m}$, the similarity between the user's preferences and barriers and the message, $\cos(\mathbf{a}, \mathbf{b})$, as defined in (10), the absolute SoC difference

[10] a spatial representation of text [22]

[11] Please note that the weighting vector $\boldsymbol{\beta}_{\mathrm{BCT},s}$ from (4) is not applied here since no empirically determined factors are used for the ML-based algorithms. Therefore, vector $\mathbf{b}$ as defined in (3) is used directly instead of vectors $\mathbf{u}_s$.

between the user and the message $|S_u - S_m|$ as in (1), again without empirically determined weighting factor, and a keyword matching between the user's preference/barriers and the message $K_{\mathrm{u,m}}$ as defined in (7):

$$\mathbf{x}_{u,m} = \left[\mathbf{b}^T \cdot \mathbf{m}, \cos(\mathbf{a}, \mathbf{b}), |S_u - S_m|, K_{\mathrm{u,m}}\right]^T \tag{12}$$

A vector $\mathbf{x}'$ of length 7×1 is defined combining the user features $\mathbf{x}_u$ from (11), and user-message relationships $\mathbf{x}_{u,m}$ from (12).

$$\mathbf{x}' = \left[\mathbf{x}_u^T, \mathbf{x}_{u,m}^T\right]^T \tag{13}$$

Naïve Bayes (NB) Classifiers: A NB classifier is commonly in RS [22]. Naïve Bayes assumes that the probability that an event occurs (e.g., user i rated item j with 4 stars out of 5) is independently influenced by all the selected input variables X [22].

Support Vector Machines: SVMs are supervised ML models used for classification (support vector classification (SVC)) and regression (support vector regression (SVR)) [26]. For SVC, the goal is to find the decision boundary which can assign most of the samples to the correct class.

K-Nearest Neighbour Classifiers: The nearest neighbours method looks for the closest neighbours w.r.t. the (Eucledian) distance in the feature space between a new sample and the training samples [26]. The number of closest neighbours K is a parameter of the KNN approach [26].

2.4 Algorithm assessment

The algorithms' performance will be assessed in terms of typical metrics *accuracy* A and *precision* P which can be derived from the confusion matrix (CM), shown in Table 2 for the case of binary classification.

Table 2. Confusion matrix for a binary classification task. TP: true positive, TN: true negative, FP: false positive, FN: false negative.

		Predicted		
		0 (negative)	1 (positive)	Total
True	0 (negative)	TN	FP	TN+FP
	1 (positive)	FN	TP	TP+FN
	Total	TN+FN	TP+FP	N

Each row in the CM in Table 2 corresponds to the instances in a predicted class, while each column represents the instances in an actual class. The *precision*

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

assesses the model's ability to correctly identify true positive instances out of all instances predicted as positive. The precision metric is particularly useful when the cost of false positives is high or when we want to ensure that positive predictions are highly reliable. *Accuracy*

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{15}$$

is a measure of the overall correctness of the model predictions, regardless of the class.

3 Study design

3.1 Enrollment

Flyers (in PDF format) were shared through the main researcher's WhatsApp and WeChat groups. Additionally, printed flyers were posted within the faculties of the University of Sheffield (UoS). The inclusion criteria were as follows:

1. Healthy individuals of age between 18 and 60 years
2. People studying or working at UoS who can read English
3. People with access to the internet through a computer, phone, or tablet

Participants were required to read the online information sheet and sign the electronic consent form before joining the study.

3.2 Experiments

Experiments were divided into two steps: a baseline assessment and daily message ratings for a duration of 7 days. After participants enroled, a baseline assessment was carried out to collect participants' geographic information and physical activity-related features for the RS. An account was created for each participant including a unique URL to collect their ratings. Participants were encouraged to rate at least three messages per day. On the rating start date, each participant received two messages through his preferred communicational channel (Email, WhatsApp, SMS or WeChat). The first one was a thank you letter, with a brief introduction to the study. The second was a reminder:

> **Dear "name",**
> **Thanks for your great support!**
> **This is the "x" day for this trial. Click url to check your new messages and provide your ratings.**
> **If you want to set another time to receive this reminder, simply reply to this message.**
>
> **NOTE: Sometimes the server might be slow, please refresh your web browser if it takes a longer time than expected.**

The reminder was sent daily at 11:00 am for seven days; the name and URL were adapted according to each participant's chosen nickname in the baseline assessment, and the "x" was the countdown day of the study. Four-dimensional ratings were requested for each message (as shown in Fig. 1 in the 3^{rd} panel): (perceived) *usefulness*[12], (perceived) *relevancy*[13], (perceived) *enjoyment*[14] and (perceived) *active trust*[15]. These four dimensions were drawn from the conceptual model [29] and adapted to the CTHC.

3.3 Web-based message delivery

Fig. 1 shows four interface pages of the designed web-based application for message delivery, which users could access using their personal URL provided in the daily reminder. It comprises:

1. a *homepage*, where users find information related to the project, an introduction and information on how to use this web application for rating messages;
2. a *list of new messages*, where users find all titles of unrated messages. The unrated messages were randomized and displayed as a list on this page for the user to view;
3. *message content* and *rating*, where users read the message and rate it;
4. a *list of rated messages*, where users can find the messages they have rated.

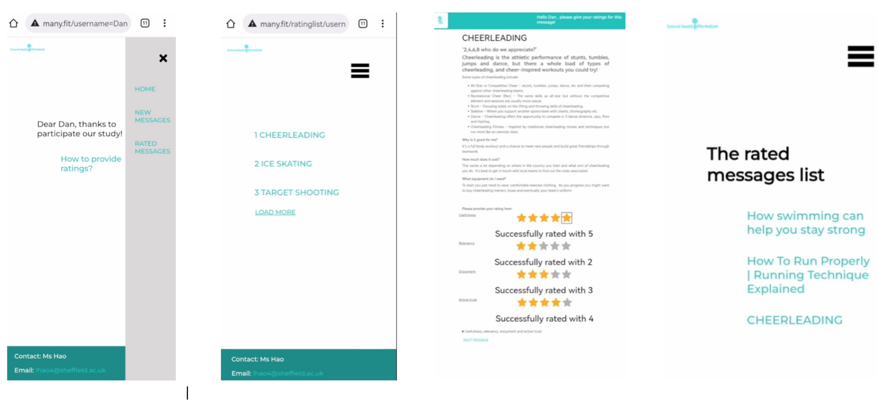

Fig. 1. Webpage layout, home page (leftmost), rating list, message content, and rated message list (rightmost)

[12] The degrees of the user's belief that the information is beneficial. *Usefulness* is the cognitive perceptions of *Efficiency* and *Effectiveness*.
[13] The quality of the information to the users that can be effectively used by the user.
[14] Affective perceptions on reading the message.
[15] Believe and have the confidence to act on the information presented.

4 Results

4.1 Participant Disposition

In total, 25 participants signed the consent form and completed the baseline assessment. Out of these, 15 participants rated at least one message during the 7-day study period. Table 3 presents the participants' characteristics based on the results of the baseline assessment.

Table 3. Summary of baseline assessment; participant characteristics (N=15).

Characteristics	Sub-category	No. of participants
Age	18-24	3
	25-34	11
	35-44	1
Gender	Female	8
	Male	7
SoC	SoC-1	0
	SoC-2	1
	SoC-3	1
	SoC-4	8
	SoC-5	5
Self-efficacy regarding PA (10-40)	low (10-20)	3
	medium (21-30)	6
	high (above 30)	6
Currently Smoking	Yes	0
	No	15
Reaching recommended PA level	Yes	3
	No	12
Message delivery channel	Email	13
	SMS	1
	Wechat	1

During the study, participants were asked to comment on the technical design shortcomings of the system. Regarding this, participants reported by webpage layout issues. For instance, the star rating system did not display correctly on some mobile phone models or a comment section under each message was requested. In some cases, the 'next message' button did not function properly, and the website occasionally reported errors during use. All reported design and technical issues were considered minor and resolved immediately during the study and therefore should not influence the results reported in the following.

4.2 Technical evaluation

In total, $1,102$ rating records were collected from the participants. Fig. 2 shows the number of rating records collected from each participant.

Given the high correlations observed among the four-dimensional ratings (see Table 4), the "average" of these ratings, referred to as the *average* in the following, is calculated by aggregating the values across the four dimensions.

As the decision rules and persuasive messages in the system were designed for participants with SoC below 5 ($S_u \leq 4$), records collected from participants

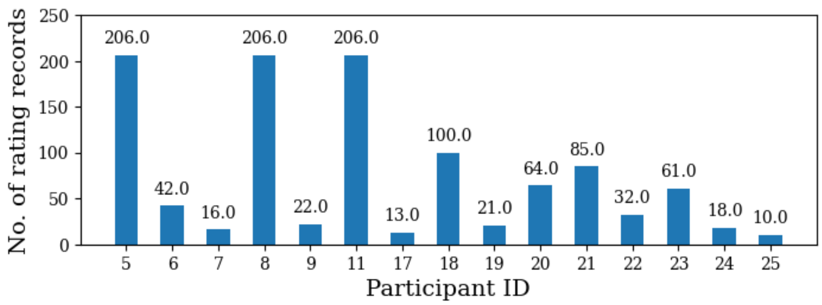

Fig. 2. Number of ratings per participant (participants without ratings are omitted).

Table 4. Correlations between 4-dimensional ratings

	Usefulness	Relevancy	Enjoyment	Active trust
Usefulness	1	0.83	0.82	0.81
Relevancy	0.83	1	0.79	0.83
Enjoyment	0.82	0.79	1	0.79
Active trust	0.81	0.83	0.79	1

with their SoC level equal to or above 5 ($S_u > 4$) were excluded to avoid bias in evaluating the system performance. Therefore, 627 records remained for evaluating the RS system. The obtained rating scores, as shown in Fig. 3, with ratings between 1 and 5 To converte the dataset to Boolean values, the Boolean record was set to 1 if the participant's rating was equal to or above 4, otherwise, it was set to 0 (see. Fig. 4).

Rule-based Algorithm: CMs were calculated to assess the performance of the RS as shown in Table 5. For the rule-based RS, precision on 1, as defined in (14), is 0.68 for *usefulness*, 0.65 for *relevancy*, 0.54 for *enjoyment*, 0.66 for *active trust*, and 0.60 for averaging all 4 dimensions. See also Table 6 for accuracy results.

Table 5. Confusion matrices of rule-based algorithm

(a) Usefulness

		predicted 0	predicted 1
true	0	81	40
true	1	107	86

(b) Relevancy

		predicted 0	predicted 1
true	0	84	46
true	1	104	80

(c) Enjoyment

		predicted 0	predicted 1
true	0	95	58
true	1	93	68

(d) Active trust

		predicted 0	predicted 1
true	0	80	43
true	1	108	83

(e) Average

		predicted 0	predicted 1
true	0	91	51
true	1	97	75

ML-based Algorithms: For the ML-based RSs, different combinations of the participants' attributions and messages' attributions were evaluated as the independent input features as listed in the 3$^{\text{rd}}$ column of Table 6, i.e. combinations

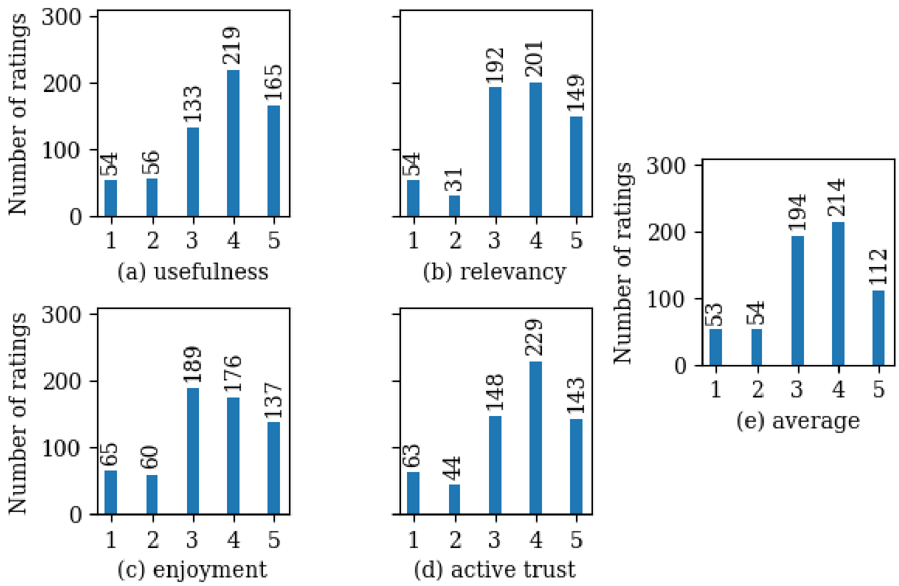

Fig. 3. Rating distribution for categories (top left to bottom middle) (a) *usefulness*, (b) *relevancy*, (c) *enjoyment*, and (d) *active trust*, as well as (e) averaged ratings.

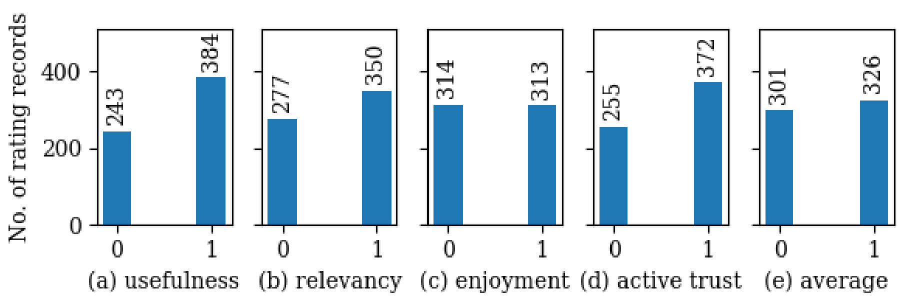

Fig. 4. Rating records distribution Boolean type 0 and 1.

from (11), (12), (13) and (10). ML algorithms include KNN KNeighbors classifier and a variant, i.e. NearestCentroid, Gaussian Naïve Bayes, linear and non-linear SVCs [26], as listed in the 1^{st} and 2^{nd} columns of Table 6.

For ML-based RS, the highest precision values for 1 were as follows: 0.76 (*Usefulness*, generated by SVM non-linear); 0.84 (*Relevancy*, generated by KNNeighbors Classifier); 0.85 (*Enjoyment*, generated by KNNeighbors Classifier); 0.77 (Active Trust, generated by Gaussian Naive Bayes); and 0.74 (Average, generated by KNeighbors Classifier). The highest value of accuracy were 0.72 (Usefulness, generated by KNNeighbors Classifier); 0.68 (Relevancy, generated by KNNeighbors Classifier ad Gaussian Naive Bayes); 0.62 (Enjoyment, gener-

Table 6. Algorithm performance in terms of accuracy A and precision P (precision on 1) depending on classifiers and input feature combinations; asterisk (*) indicates that the classifier decided on a single category for all data. Bold font indicates best performance.

Algorithm		Input feature $\mathbf{x}$	Usefulness		Relevancy		Enjoyment		Active trust		Average	
			A	P	A	P	A	P	A	P	A	P
Rule-based algorithm		$[S_{u,m}, B_{u,m}, K_{\mathrm{u,m}}]^T$	0.53	0.68	0.54	0.65	0.52	0.54	0.52	0.66	0.53	0.60
KNN	K-Nearest Neighbor Classifier	$\mathbf{x}_u$	0.66	0.68	0.55	**0.84**	**0.61**	**0.85**	**0.68**	0.70	**0.67**	**0.72**
		$[\mathbf{x}_u^T, \cos(\mathbf{a}, \mathbf{b})]^T$	**0.72**	**0.75**	**0.68**	**0.71**	0.62	0.66	**0.66**	**0.72**	**0.69**	**0.74**
		$\mathbf{x}_{u,m}$	0.65	0.68	0.59	0.65	0.51	0.52	0.60	0.67	0.59	0.62
		$\mathbf{x}'$	0.68	0.71	**0.67**	**0.74**	0.56	0.57	0.60	0.67	0.65	0.67
	K-Nearest centriod	$\mathbf{x}_u$	0.57	0.61	0.65	0.67	0.57	0.56	0.5	0.61	0.48	0.51
		$\mathbf{x}_{u,m}$	0.55	0.66	0.54	0.63	0.49	0.51	0.53	0.65	0.52	0.56
		x'	0.51	0.60	0.65	0.67	0.56	0.55	0.5	0.59	0.50	0.53
NB	Gaussian NB	$\mathbf{x}_u$	0.61*	0.62*	0.65	0.67	0.57	0.56	0.61*	0.61*	0.59	0.58
		$\mathbf{x}_{u,m}$	0.68	0.71	0.65	0.68	0.59	0.61	0.66	0.69	0.66	0.68
		$\mathbf{x}'$	0.61*	0.61*	0.68	0.76	0.58	0.57	0.52	**0.77**	0.65	0.72
SVM	SVC linear	$\mathbf{x}_u$	0.61*	0.61*	0.65	0.67	0.56	0.57	0.61*	0.61*	0.59	0.58
	SVM non-linear	$[\mathbf{x}_u^T, \cos(\mathbf{a}, \mathbf{b})]^T$	0.54	**0.75**	0.51	0.58	0.53	0.57	0.52	0.61	0.59	0.69
		$[\mathbf{x}_u^T, K_{\mathrm{u,m}}]^T$	0.59	**0.76**	0.56	0.61	0.57	0.57	0.53	0.64	0.40	0.41

ated by KNNeighbors Classifier); 0.68 (Active Trust, generated by KNNeighbors Classifier); and 0.69 (Average, generated by KNNeighbors Classifier). The results may vary slightly with different randomly split datasets; however, the KNNeighbors classifier demonstrated the best overall performance among all the RS.

Compared to randomised message delivery, the rule-based algorithm leads to a relative improvement of 20 % precision in average for positively rated messages over the four tested dimensions in comparison to random message selection. For the machine-learning-based algorithms, KNN leads to the best prediction of user ratings with a relative improvement of 48 % in average over the four tested dimensions, with 70 % relative improvement for dimension *enjoyment*.

5 Conclusion

5.1 System performance on recommending tailored health messages

In conclusion, the obtained results show that the developed system is feasible for use as a recommendation system for tailoring messages related to PA promotion. The rule-based algorithm is superior to random message delivery, and it can be used in the initial stage of message selection when no historical data from users is available for training the ML-based algorithms. With limited data, the KNN method, using user features and text similarity scores, results in the best performance among the tested algorithms and combinations of independent variables (input features), markedly improving overall performance. Therefore, when dealing with a small amount of collected data, KNN can improve the performance of health message RS. The results of the correlation test reveal a significant correlation among the 4-dimensional ratings, which is in line with findings by Crutzen (2011) that active trust can act as a mediator for user perceptions. These perceptions encompass effectiveness (usefulness and relevancy) and enjoyment, ultimately leading to e-loyalty. [29].

5.2 Effective independent variables

In this study, participants' characteristics were selected to generate user-message relationships. The participants' SoC, self-efficacy on PA, the current status of PA level, and text similarity between participants' preferences/barriers with messages were shown to be effective in increasing the precision rate in selecting the right messages for participants. It was shown that, using the rule-based system, a relationship between each participant and each message (matching score) can be established based on behaviour theories and BCT. The participants' SoC and current status of PA level can be linked to the messages' content by comparing with the messages' best-fitted SoC and PA level. BCTs necessary for behaviour change of participants can be linked to the messages' BCTs taxonomy [14]. Participants' preferences in general and barriers to performing PA can be used to link messages' content through keyword matching and text similarity matching. Due to current behaviour change theories providing more frameworks (i.e. what should be considered) than quantitative intervention doses (i.e. how much should be considered), the rule-based system still needs to identify the right coefficient (β's) for each independent variable to increase accuracy and precision. In the proposed RS based on ML algorithms, different classifiers achieved similar good performance with different combinations of independent variables (input features). Participants' features and text similarity between participants and messages worked better in the KNNeighbors Classifier and the non-linear SVM. The difference of SoC, number of BCTs matching, and text similarity between the participants and the messages worked better with the Gaussian Naïve Bayes classifier. In future studies, the rule-based system could be used as long as there are insufficient annotations (ratings). When the system obtains more than approx. 600 ratings from the participants, the KNNeighbors Classifier can be trained to generate new recommendations.

5.3 Limitations

The main limitation of this study is the limited size of the dataset. As a pre-pilot study, it aimed to assess the feasibility of the proposed approach. However, limited participants and a short study period resulted in a small dataset. The second limitation is the cohort of participants. The system used in this study is intended for people with COPD, but the participants did not have COPD, leading to the omission of some independent variables (e.g., COPD level, smoking status). This may contribute to low accuracy in message selection. Although the system achieved moderate precision in recommending participants' preferred messages, the accuracy was still relatively low, causing a significant loss of participants' preferred messages. To increase accuracy, future studies are needed to determine the coefficient of independent variables and to improve the algorithms.

5.4 Discussion

In addition to the advantages a RS for CTHC in obtaining more information related to physical health and psychological status, ML algorithms can further enhance its effectiveness. By integrating behaviour change theories and machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Gaussian Naive Bayes (Gaussian NB), a more personalized and effective health communication strategy can be created.

Unlike the industry field (content provide such as Youtube and Netflix, E-commerce company such as Amazon and Alibaba), one of the advantages of CTHC with a RS is the ability to obtain more information related to physical health and psychological status. This is particularly beneficial in addressing the challenge of the 'cold start' problem faced in the industry field, where it becomes difficult to recommend accurate information due to the presence of new users or limited data in the system. The obtained information can be integrated with behaviour change theories and models, along with ML algorithms, to create a more personalized and effective health communication strategy.

CTHC faces the challenge of categorising tailored information for users to increase adherence in the long term and maintain continuous inner motivation for behaviour change. Breaking down persuasive health information into smaller elements based on behaviour change theories and BCT holds promise, but the quantitative effectiveness of these elements remains unknown. This presents a significant challenge for current CTHC systems to consistently improve the effectiveness of behaviour change interventions when targeting different populations and behaviours.

To address this challenge, future studies should focus on investigating the correlations among user characteristics, health information features, user preferences on health information, and the effectiveness of behaviour changes. By exploring these correlations, researchers can optimise and refine CTHC strategies, leading to more successful and personalised behaviour change interventions.

These efforts to integrate machine learning algorithms and conduct further research would contribute to the continuous improvement of CTHC, ultimately enabling more successful behaviour change interventions tailored to individual needs.

Bibliography

[1] World Health Organisation: Physical activity, online: https://www.who.int/news-room/fact-sheets/detail/physical-activity (accessed: 31.03.2023)

[2] Gill, D., Hammond, C., Reifsteck, E., Jehu, C., Williams, R., Adams, M., Lange, E., Becofsky, K., Rodriguez, E., Shang, Y.: Physical activity and quality of life. J Prev Med Public Health (2013)

[3] Op den Akker, H., Cabrita, M., Op den Akker, R., Jones, V.M., Hermens, H.J.: Tailored motivational message generation: A model and practical framework for real-time physical activity coaching. Journal of Biomedical Informatics 55, 104–115

[4] Grant.Harrington, N., M.Noar, S.: Reporting standards for studies of tailored interventions. Health Education Research 27(2), 331–342 (Apr 2012)

[5] Lustria, M.L.A., Cortese, J., Noar, S.M., Glueckauf, R.L.: Computer-tailored health interventions delivered over the web: Review and analysis of key components. Patient Education and Counseling 74(2), 156–173 (2009)

[6] Noar, S.M., Harrington, N.G.: Tailored communications for health-related decision-making and behavior change. In: Diefenbach, M.A., Miller-Halegoua, S., Bowen, D.J. (eds.) Handbook of Health Decision Science, pp. 251–263. Springer

[7] American Psychological Association: Stage theory, https://dictionary.apa.org/stage-theory, (accessed: 05.04.2023)

[8] Sadasivam, R.S., Borglund, E.M., Adams, R., Marlin, B.M., Houston, T.K.: Impact of a collective intelligence tailored messaging system on smoking cessation: The perspect randomized experiment. J. Med. Internet Res. 18(11)

[9] Sutton, S.: Health behavior: Psychosocial theories. International Encyclopedia of the Social & Behavioral Sciences pp. 6499–6506 (2001)

[10] Noar, S.M., Benac, C.N., Harris, M.S.: Does tailoring matter? meta-analytic review of tailored print health behavior change interventions. Psychological Bulletin 133(4), 673–693 (2007)

[11] Prochaska, J.O., Velicer, W.F., Rossi, J.S., Goldstein, M.G., Marcus, B.H., Rakowski, W., Fiore, C., Harlow, L.L., Redding, C.A., Rosenbloom, D., Rossi, S.R.: Stages of change and decisional balance for 12 problem behaviors. Health psychology (1994)

[12] Glanz, K., Rimer, B.K., Viswanath, K.: Health Behaviour and health education: theory, research, and practice. A Wiley Imprint (2008)

[13] Ajzen, I.: The theory of planned behavior. Organizational Behavior and Human Decision Processes 50(2), 179–211 (1991)

[14] Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M.P., Cane, J., Wood, C.E.: Bct taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. Annals of Behavioral Medicine 46, 81–95 (Aug 2013)

[15] Noar, S.M., Webb, E.M., Stee, S.K.V., Redding, C.A., Feist-Price, S., Crosby, R., Troutman, A.: Using computer technology for hiv prevention among african-americans: development of a tailored information program for safer sex (tipss). Health Education Research 26(3), 393–406 (2011)

[16] Zhou, M., Fukuoka, Y., Mintz, Y., Goldberg, K., Kaminsky, P., Flowers, E., Aswani, A.: Evaluating machine learning–based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial. JMIR Mhealth Uhealth 6(1)

[17] Ryan, N.M., O'Donovan, C.B., Forster, H., Woolhead, C., Walsh, M.: New tools for personalised nutrition: The food4me project. Nutrition Bulletin 40(2), 134–139 (2015)

[18] Franco, R.Z., Fallaize, R., Hwang, F., Lovegrove, J.A.: Strategies for online personalised nutrition advice employed in the development of the eNutri web app. Proceedings of the Nutrition Society 78(3), 407–417

[19] Lustria, M.L.A., Noar, S.M., Cortese, J., Stee, S.K.V., Glueckauf, R.L., Lee, J.: A meta-analysis of web-delivered tailored health behavior change interventions. Journal of Health Communication 18(9), 1039–1069 (2013)

[20] Hao, L., Goetze, S., Alessa, T., Hawley, M.: Effectiveness of computer tailored health communication in increasing physical activity in people with or at risk of long-term conditions: systematic review and meta-analysis. Subm. to J. of Medical Int. Res. (JMIR) (2023)

[21] Sadasivam, R.S., Cutrona, S.L., Kinney, R.L., Marlin, B.M., Mazor, K.M., Lemon, S.C., Houston, T.K.: Collective-Intelligence Recommender Systems: Advancing Computer Tailoring for Health Behavior Change Into the 21st Century. Journal of Medical Internet Research 18(3) (Mar 2016)

[22] Ricci, F., Rokach, L., Bracha, S.: Recommender Systems Handbook. Springer US (2015)

[23] Cheung, K.L., Durusu, D., Sui, X., de Vries, H.: How recommender systems could support and enhance computer-tailored digital health programs: A scoping review. DIGITAL HEALTH 5

[24] Schwarzer, R.: Self-Efficacy - Thought Control Of Action. Taylor & Francis (1992)

[25] Schwarzer, R.: The Health Action Process Approach (HAPA), `https://www.hapa-model.de/`, (accessed: 01.12.2022)

[26] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research 12(Oct), 2825–2830 (2011)

[27] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Conf. Learning Representations (2013)

[28] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT (2019)

[29] Crutzen, R., Cyr, D., de Vries, N.K.: Bringing loyalty to e-health: Theory validation using three internet-delivered interventions. J Med Internet Res (2011)