Recommending Activities for Mental Health and Well-being: Insights from Two User Studies

Rohani, Darius Adam; Springer, Aaron; Hollis, Victoria; Bardram, Jakob Eyvind; Whittaker, Steve

Published in:
IEEE Transactions on Emerging Topics in Computing

Link to article, DOI:
10.1109/tetc.2020.2972007

Publication date:
2021

Document Version
Peer reviewed version

Citation (APA):
Recommending Activities for Mental Health and Well-being: Insights from Two User Studies

Darius A. Rohani, Member, IEEE, Aaron Springer, Victoria Hollis, Jakob E. Bardram, Member, IEEE, and Steve Whittaker,

Abstract—Engaging in daily activities that engender positive affect (e.g., exercise and socializing) is critical for emotional well-being and is effective in reducing clinical depression. However current digital mental health interventions have not exploited recommender approaches to encourage such healthy behaviors. This paper tests the feasibility of recommending personalized healthy activities to users. Using two mobile applications, we collected high-quality data about specific healthy activities from two populations: a clinical sample diagnosed with a mood disorder (n = 318 activities/user) and a non-clinical sample (n = 59 activities/user). Activities were labeled with a type (e.g., social, leisure, work) and rated for their impact on mood. We used a probabilistic multinomial Naive Bayes (NB) Classifier and a Support Vector Machine (SVM) to model the activities as a bag-of-words to predict mood outcome. We separate the analysis into a generalized model where we pooled all participants, comparing it with a personalized model. In both the clinical and non-clinical samples there was a significant difference between the two models. Both NB and SVM favored the personalized model after collecting 58.92 (SD = 20.96) activities. This research sheds light on recommendations for mental health, showing that personalization is key for recommending the right activity to each user.

Index Terms—Recommendation, Activity, Depression, Well-being, Behavior, Content-analysis, MHealth, Naive Bayes

1 INTRODUCTION

Mental health and well-being are major public and personal health issues, being a leading cause of disability and disease burden worldwide [1]. 30% of men and 40% of women experience a major depressive episode at least once in their life with minorities being even more vulnerable [2]. Depression is second only to heart disease as a cause of disease burden [3], and many people remain untreated [4]. As a result, there is a growing research effort to develop personal informatics technologies to alleviate mental health symptoms (see [5], [6] for reviews). These have resulted in innovative mobile (mHealth) solutions [7], [8], [9], [10], [11], [12]. Nevertheless, there are important limitations to these existing applications; they often rely on burdensome manual user tracking of symptoms, and intervention strategies are sometimes based around theoretically unmotivated self-help heuristics. Together these limitations may explain relatively low levels of long-term adoption [13], [14]. The current situation parallels the early era of recommender platforms, which were also driven by simple heuristics. Now, of course, many recommender platforms reduce the analytic burden for users by providing explicit product recommendations, such as books, movies or songs. This not only increases platform usage [15], but also improves user satisfaction [16].

This paper explores the application of recommender techniques to the mental health domain. The key scientific finding that motivates our work is that when patients regularly enact activities that they rate as pleasant, this has significant positive impacts on mental health [17], [18]. A successful behavioral intervention therefore relies on recommending personal activities that are likely to improve the patient’s affective state. For example, walking the dog or meeting friends commonly boost mood. This simple Pleasant Event Scheduling (PES) intervention has been demonstrated to be an effective treatment for depression, with effects that are comparable to standard psychotherapy methods such as Cognitive Behavioral Therapy (CBT) treatment and drug therapies [19], [20].

However, most prior clinical PES interventions rely on non-digital approaches where participants manually log their daily behaviors for several weeks. Patients then collaborate with their therapist to analyze this record, to identify ‘pleasant’ events. Patient and therapist then generate a treatment plan involving a weekly schedule of those pleasant events. However, there are drawbacks to this approach. First, the process of manually identifying pleasant events may be prone to errors, as the effect of a given event can be specific to an individual. There is also time and cost involved in scheduling therapist meetings. As a result, patients are receptive to computational approaches: with surveys showing 91% of patient respondents are positive about computationally enabled therapies because of the anonymity and control they offer [21]. An app can be accessed privately without needing to disclose personal details or schedule official health services [22].

The goal of our current research is therefore to explore the viability of recommendation methods in the context of PES. Specifically, we examine whether recommender methods might allow us to reliably predict such mood-boosting
events. Generating reliable predictions would allow a future system to use these models to recommend positive activities to users.

This study uses data collected from two mobile applications that allow clinical and non-clinical participants to manually log details of personal activities along with mood ratings. We assess whether positive ratings for different activity types can be successfully predicted using automatic methods. We also explore individual differences to determine whether personalized versus general models best predict ratings. If successful predictions can be achieved, these recommendations could be incorporated into well-being apps, enabling recommendations about future activities which – according to PES theory – could improve a user’s mental health.

Our main contributions are:

1) Offline analysis of logged activities from patients with a clinical depression and non-clinical users to test the feasibility of predicting activities that are positive for individual users.

2) We ran holdout cross-validation with incremental training size on the logged activities to show the relation between a personalized or general model as a function of the number of activities registered.

3) We performed follow-up evaluations on the trained recommender models to extract key insights about the activity parameters that had an effect on the level of positivity. Furthermore, we make the NB trained general model now publicly available allowing others to build on these insights.

4) Finally, we ran the personalized recommender models through a 4-week app usage simulation to illustrate a novel scenario of applying activity recommendation to technology for mental health.

The rest of the paper is structured as follows: Section 2 provides background information on current recommender systems for mHealth, and a summary of the different types of recommender models that utilize textual information. Section 3 describes the user studies that we conducted to compile our database of activities, including a detailed description of how we analyzed the activity data. Section 4 presents the results of our recommender model which predicts activities that were rated positively by participants. The section also derives insights from the models, indicating why a given activity was classified as positive. The section ends with a simulation study to demonstrate how the model could provide novel activity recommendations in a future study. Section 5 discusses our findings and implications for future work. Section 6 concludes our study.

2 Recommender systems in m-Health

In this section we review relevant work on recommender systems for health and well-being, and on designing recommender systems for textual knowledge.

2.1 Health and Well-being Recommendation

Many recent systems deploy smartphones to improve mental health, well-being, and life quality (see [23], [24], [25], [26], for reviews). However, few of these systems have incorporated recommender approaches. Wahle et al. [27] developed the Mobile Sensing and Support system (MOSS), providing context-aware activity recommendations to alleviate depressive symptoms. They used context features such as time-at-home, number of steps, and number of phone calls to derive a ‘basket’ score within four predefined activity categories (social activity, mindfulness, relaxation, and physical activity). When users selected a target activity category they were presented with the three ‘top’ interventions, chosen based on previous 1-5 star manual ratings from the user. A single-arm deployment showed a significant drop in depressive symptoms over an 8-week period when these manual recommendations were applied. Hollis et al. [12], [28] developed an activity tracking app with a mood forecasting module. They used historical mood data to predict tomorrow’s mood, developing a regression model relating mood to different activity categories. The model was used to recommend activity categories (e.g. social events, exercise) that could potentially improve mood. Yang et al. [29] developed an emotional health app system that used an SVM and decision trees to classify users’ depressive level. Based on the depression stage, different pre-specified actions were recommended, e.g. music therapy or suggestions to plan social activities.

These last two systems both show promising results and the latter showed validity, but neither focused on recommending specific positive activities. This is a significant weakness as identifying a specific activity significantly increases the probability of users enacting that activity [30], [31]. While Wahle et al. [27] do suggest specific activities, they do not leverage information about those activities but rather rely on simple past user activity ratings. This approach does not follow standard recommendations methodology such as item-by-item collaborative filtering [32].

2.2 Textual Knowledge

Raw text is almost always present in recommender scenarios, whether it is songs, books, movies, or restaurants. But some recommender systems do not directly link that textual information to the target item [33]. Often similarities in ratings among users or between items are used to recommend items using collaborative filtering [32]. Mining textual information – as sometimes seen in content-based recommender systems – has the advantage of extracting knowledge from unrated items to build a personalized model that fits users with unique, individual tastes [34]. Although no prior work has looked specifically at text directly associated with personal activities, several systems leverage other aspects of textual knowledge. Kim et al. [35] used a Recurrent Neural Network to extract textual knowledge from the titles of music playlists. By supplementing a collaborative filtering recommender system with textual knowledge, they were able to improve music playlist recommendations on all evaluation parameters.

Faridani [36] used a supervised learning method called Canonical Correlation Analysis to model the relationship between textual descriptions (e.g., a hotel review) and the corresponding numerical rating. This method projects words into a high-dimensional space and uses a distance measure to find correlations between input texts. Mocherla
et al. [37] used a Support Vector Machine (SVM) on Wikipedia text. It also projects the text into a high-dimensional space (each word corresponding to a dimension).

While the above models show good performance, they all rely on an extensive collection of known data (training sets) to fit model parameters. A simpler non-parametric Naive Bayes (NB) method is still one of the most popular solutions for leveraging textual knowledge [38]. Besides being non-parametric it is computationally efficient and scales well. Mocherla et al. [37] also compare SVM with NB, showing a reduction in training time from 17 to 4 hours on the full Wikipedia corpus; the only major performance decrement was in recommending topics that overlapped with others.

Textual deep learning analysis such as Recurrent Neural Networks of e.g., movies and songs are possible for large databases containing millions of samples. However no equivalent large database of end-user activity data exists. We ran two in-the-wild user studies for a prolonged period to build a catalog of thousands of activities. As we are dealing with a vulnerable target group we chose to use a recommender model with fast convergence i.e., showing good performance based on a limited set of recorded activity data. A further requirement for our target application is that the recommendation is transparent. For a user to be motivated to enact a recommended activity they must be able to determine why an activity is recommended. Furthermore, the model has to be efficient and lightweight as it has to operate in a mobile context. For all the above reasons we chose not to explore deep learning models as they require larger datasets, feature several trainable weights that are hard to interpret and require more computational power. Instead we chose to model our data using NB as a fast converging, and transparent model that fulfills our requirements for a mHealth recommender system on textual activity data. We included the SVM to benchmark it against NB, as SVM is known for its better performance [37].

### 3 Method

As outlined above, recommender systems for well-being are at an early stage, especially for mental health. Therefore, there is no pre-existing dataset for evaluating recommender model performance. Such a dataset would require detailed activity registration and labeling over an extended period of time, along with detailed information about patient profiles.

In this study, we therefore utilize two datasets from two smartphone user studies, which collected data ‘in-the-wild’ for 3 and 4 weeks respectively. These apps generated high-quality data from participants who logged their activities together with labels for activity type and the immediate effects of that activity on mood or pleasure. The dataset possesses two key advantages over typical user-behavior modeling datasets: 1) users provide real-time longitudinal self-generated labels for all attributes without reliance on crowd-sourced annotations, and 2) users log their own perceptions of how activities influence mood and also include textual information, producing a reliable dataset.

Both studies have been approved by the relevant Ethical Committees and Internal Review Boards (i.e., 17018289 and HS2466, respectively). Given the sensitive and personal nature of the logged data, these approvals required the researchers to sign a confidentiality agreement with the participants stating that their logged activity data would be kept confidential.

The two user studies targeted two distinct user groups; the first involved patients with a clinical diagnosis of unipolar (i.e., depression) or bipolar disorder, while the other was a non-clinical sample. In total, the two datasets respectively consist of 1,684 entries from 7 patients, and 6,344 entries from 134 users.

#### 3.1 Clinical Sample

For the clinical sample, we collected data via the MORIBUS app [39]. It is designed to support BA [40] in which patients log hourly activity information each day, rating each activity with a ‘pleasure’ score, i.e. the extent to which the activity engenders positive or negative emotion. This detailed information is later used by a therapist to identify activities that influence the depressive symptoms of the patient. Together, the patient and the therapist plan the following week’s activities based on recommendations from the therapist, and schedule positive activities that could lead to healthy behavior change. To replicate the BA procedure in the MORIBUS app, each patient was instructed to plan a set of activities for a day or week and evaluate these activities on an hourly basis. This is done using a daily calendar with hour-based timeslots from 8 am (08:00) to 10 pm (22:00). An activity is created and planned using the activity registration page shown in Figure 1 (left). When creating an activity, patients: (i) select an activity type (e.g., Leisure) as adopted by [41], and (ii) choose from a precompiled list of typical activities of the chosen type (by pressing the light bulb button), or manually enter an activity. When an activity was marked as done, patients were notified to enter a positivity rating, i.e., how much they enjoyed the activity. This is shown in Figure 1 (right). Examples of collected data are shown in Table 1.

#### 3.2 Non-clinical Sample

For the non-clinical sample, we collected data using the EmotiCal app [12, 28]. EmotiCal is a mobile application created to help regular people manage their mood and improve well-being using PES. EmotiCal users logged two daily entries in which they manually recorded mood, energy
level, and activity labels. For example, users can log social interactions (e.g., time spent with a friend or coworker), aspects of physical health (e.g., sleep or exercise), or work activities (e.g., meetings) to track these activities’ effects on mood. EmotiCal also prompts users to generate short textual explanations of how and why they think those activities affected mood.

Data was collected from EmotiCal using a slightly different procedure, in which participants first rated their current mood and then assessed which of their recent activities had induced that mood. To generate Mood ratings, participants first reported their current mood on a scale from -3 (very negative) to +3 (very positive). These ratings were generated against a standardized scale where the different scalar values were clearly defined as part of the app registration process. Participants next reported and classified recent activities that contributed to that rating, and finally generated a textual description of those activities. A screenshot indicating how the participants logged relations between Mood and activities is shown in Figure 2. Mood rating is shown on the left side, and activity classification on the right side. Users chose the activities that contribute to their current mood. Participants also rated Energy levels (left hand side), but we did not use these ratings in the current analysis. Examples of EmotiCal textual descriptions along with their activity labels and mood ratings are shown in Table 2.

3.3 Data Processing

The activity data collected from the two studies were analyzed in MATLAB (vs. R2018B). The processing steps are outlined below:

- In the clinical study, activity entries from Danish-speaking patients were translated by the authors into English.

- A custom script pre-processed textual data into (i) lower case and (ii) removed all special characters, dots, commas, trailing and leading spaces from the activity text. Then the dimensionality of unique words was reduced further by (iii) removing typical words determined by a list of 665 stop-words, (iv) Porter stemming words (to combine words such as “cleaned” and “clean”) and (v) removing long words ($n > 15$ characters).

- We removed data entries that had empty text strings.

- Data from users who provided fewer than 3 different mood ratings were removed due to low variability in the outcome measure.

- The ratings were transformed to a binary outcome.

---

**TABLE 2**

Example of data collected via the EmotiCal app.

<table>
<thead>
<tr>
<th>ID</th>
<th>Time</th>
<th>Activity</th>
<th>Label</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>08:23</td>
<td>I made time to do things I like, watch TV and chat with a friend. So it’s been a good day</td>
<td>Leisure, Social</td>
<td>2</td>
</tr>
<tr>
<td>S8</td>
<td>08:27</td>
<td>Had delicious chicken dinner with mashed potatoes and broccoli tonight. Will head to amusement park tonight with my family</td>
<td>Food, Leisure</td>
<td>1</td>
</tr>
<tr>
<td>S17</td>
<td>08:08</td>
<td>I get to be with my boyfriend</td>
<td>Social</td>
<td>1</td>
</tr>
<tr>
<td>S42</td>
<td>08:04</td>
<td>Drinking homebrew with an astronaut</td>
<td>Work, Leisure, Social</td>
<td>2</td>
</tr>
<tr>
<td>S51</td>
<td>08:03</td>
<td>Driving home in traffic - not my favorite thing to do</td>
<td>Work</td>
<td>-1</td>
</tr>
</tbody>
</table>
variable (C) indicating a positive activity or otherwise. This was done for each user separately, where the median was used as a cut-off. We transformed the outcome variable to a binary variable to be consistent with existing approaches taken for movie, song, and book recommendations.

- Finally we transformed the activity text and labels to bag-of-words.

To assess the degree of personalization achieved by the trained recommender models, we ran individual subjects models through a public dataset of 320 pleasant activities [17]. An independent researcher manually applied activity labels to all activities in that dataset. Furthermore, we evaluated the impact of two features 'activity text' and 'label' by assessing their individual information gain on the prior probability through a Kullback-Leibler divergence analysis.

### 3.4 Statistical Analysis

We used the specific activity text ($A_1$), and the labeled activity types ($A_2$) as the two features for the activity ($D$). By (i) assuming that the position of the words in the activities and the labels selected do not matter, and (ii) that they are conditionally independent of one another, we can transform the features into a bag-of-words ($d_D$), and use the naive assumption on the Bayes theorem. The multinomial conditional probability model is then written as

$$P(C_j|D) = \frac{P(C_j) \prod^{f}_{i=1} P(A_i|C_j)}{P(A_1, A_2)} \quad (1)$$

The prior probability $P(C_j)$ can be calculated as the number of activities classified as $j$ divided by the total number of activities. With the bag-of-words method the likelihood function can be calculated as:

$$P(A_i|C_j) = \prod_{k=1}^{[d_D]} P(w_k|C_j) = \prod_{k=1}^{[d_D]} \frac{f(w_k, C_j) + 1}{\sum_{t \in V} f(t, C_j) + |V|} \quad (2)$$

Where $w_k$ is the $k^{th}$ word in the bag of words $d_D$ for $A_1$. $f(x)$ is a function that counts all the number of times that the word in the training documents belonged to class $C_j$. Laplace smoothing has been added to the equation to avoid zero probabilities for unknown cases. The evidence parameter in the denominator of Eq. 1 can be omitted since it does not depend on $C$.

The SVM model followed the approach described in [37], where the frequency-inverse document frequency (tf-idf) was calculated from the bag-of-words instead of the simple count ($f(w_i, C_j)$) due to the better observed performance [37]. The standard value ($CR = 1$) was chosen for the regularization term. A sigmoid function was used to yield posterior probabilities.

### 3.5 Comparison Models

The personalized model was compared against two other cases, as listed below:

1. **Baseline**: A simplified model that classifies activities based on the most represented classes. This is equivalent to the prior probability.

2. **General**: A model that combines all participants into a single generalized dataset. All subjects' data was pooled before cross-validation.

The resulting trained generalized NB model is available as a MATLAB file (.mat) with a corresponding MATLAB script to calculate the relative conditional probability of any given activity. The files can be found on GitHub.

### 3.6 Evaluation Metrics

Model performance was evaluated by holdout cross-validation by randomly splitting each dataset into a training and a test dataset. The size of the training set was incremented from 1% to 90% to simulate a learning process when fitting the model. The data split process was repeated 20 times ($N_{rep}$) in each sample. We report mean test errors (i.e., one minus the fraction between correctly predicted cases and all cases) to take into account both the true negative and true positive cases.

Besides presenting the learning process error rate for each model we also compare the models at three different training size percentages using the F1-score, Precision and Recall metrics. Since the clinical sample had more data we evaluated the scores on smaller training sizes.

Comparison between the personalized and the general model was done through a non-parametric permutation test using the $t_{max}$ based correction for multiple comparisons (MC) [42]. We control for the MC problem by comparing the models on each sample ($c_i$), using a repeated measures $t$-statistic. The method swaps some $c_i$ values between the personalized and general results, and calculates a set of t-values. The maximum t-value across all $c_i$ is stored. This is repeated 5000 times to generate a null-distribution of maximum t-values. If the original t-values are larger or lower than 97.5% of the values in the generated null-distribution we can classify the sample point as being statistically significant different between the two models.

### 4 Results

An overview of the data collected from the two studies is shown in Table 3. The differences in population, the number

<table>
<thead>
<tr>
<th></th>
<th>Non-clinical</th>
<th>Clinical</th>
</tr>
</thead>
<tbody>
<tr>
<td># Users (% female)</td>
<td>91 (64%)</td>
<td>7 (57%)</td>
</tr>
<tr>
<td># Users after PP</td>
<td>59</td>
<td>5</td>
</tr>
<tr>
<td>Mean Age after PP</td>
<td>30.98 (SD = 10.90)</td>
<td>43.00 (SD = 19.59)</td>
</tr>
<tr>
<td>Study duration (weeks)</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td># Collected activities</td>
<td>3,921</td>
<td>1,684</td>
</tr>
<tr>
<td># Activities per user</td>
<td>3,708 (SD = 9.10)</td>
<td>318 (SD = 106)</td>
</tr>
<tr>
<td># Unique words</td>
<td>5,702</td>
<td>478</td>
</tr>
<tr>
<td># Unique words after PP</td>
<td>3,708</td>
<td>348</td>
</tr>
</tbody>
</table>


GeneralNBmodel
of activities and the method of activity reporting gave us a unique opportunity to study activity models in two different contexts. Hence, the results from the two studies are presented separately, before we generalize across the two studies.

4.1 Clinical Sample

The error rates of the recommender models, as a function of training size, are shown in Figure 3a. The shaded area represents the standard error (SE). For the personalized and baseline model, we compute the pooled SE instead to represent the SE of each user.

The NB and SVM models seem to converge and stabilize after a training size of 40%. This is verified by inspecting the test and training error (Figure 3b). Here we see low bias and variance and convergence after a training size of 100 activities (40% holdout).

We supplemented the above error analysis with the F1-score, recall, precision and AUC from the ROC-curve at three different training sizes, for a more thorough comparison between the models. The result is shown in Table 4.

We used the \( t_{\text{max}} \) based permutation to test the difference between the general and personalized model for both the NB and SVM models. The result is depicted in Figure 4. The resulting critical \( t \)-value is plotted as red horizontal lines on the resulting null-distribution in Figure 4b, 4d. Statistically significant differences between the two models are seen after a 15% holdout (45.65, SD = 16.03 activities). Examples of three holdout samples are shown on the null-distribution as grey-level colored horizontal lines.

The benefit of a personalized model was further investigated by inspecting the trained (90% holdout training size) likelihood of the activity labels. The ratio of the likelihood of either recommending an activity label or not was computed and visualized in Figure 5. A value colored in blue indicates a higher likelihood of recommending. Values colored in red indicate the opposite.

In both SVM and NB models, we see that the ‘Social’ activity type has a higher likelihood of recommendation (i.e. has a positive rating) for every subject (P1-P5), while ‘Practical’ activities do not. This confirms other studies investigating activities and well-being [41]. The other labels show personal preferences when we compare individual models. For instance, ‘Spare time’ activities are recommended for P1, but not for P3 and P4. Note too that the personal models differ from the general model for one or activity types.

To compare the information gain of activity text and type features, we ran a Kullback-Leibler divergence analysis on 90% holdout data. For the general model the information gain for the two different features was \( D_{KL}(P(C|A_1)||P(C)) = 0.017 \) and \( D_{KL}(P(C|A_2)||P(C)) = 0.001 \), suggesting that activity text carries more information regarding the variability of the posterior probability, while the opposite was true for the SVM classifier (4.81 \( \cdot 10^{-5} \) and 7.62 \( \cdot 10^{-4} \) respectively). When inspecting the information gain on a per subject level, we saw contradictory results even within NB and SVM.
### 4.2 Non-clinical Sample

Figure 6 shows the learning curve for the non-clinical sample and a representative training and test error visualization from S50. With fewer activities per subject the holdout percentage for the personalized model corresponds to a much lower number of activities (15% holdout: \( M = 8.92, SD = 1.34 \) activities). Therefore, (i) the initial error is higher, (ii) the test and training error does not converge revealing high variance, which is also the case for the clinical sample when looking at Figure 3b at around 15% holdout (60 samples), and (iii) the general model seems to outperform the personalized model early on, since we have as many as 59 subjects corresponding to 514 activities at 15% holdout.

In depth results from holdout sizes 10%, 30% and 80% are shown in Table 5. In general the personalized model improves substantially as the training size increases from 30% to 80%, compared to the general model.

The permutation test shows significant differences in all holdout points between the two models, as shown in Figure 7. The constructed null-distribution is centered more closely around zero than for the clinical sample. This is due to the larger differences seen in the early holdout samples. If only the last half of the graph was analyzed, the holdout samples after 0.7 would not have been significantly different between the models (critical t-value = 8.97).

In line with the results from the clinical study, at 90% holdout, in the NB model, we saw that ‘Social’ activity had a likelihood ratio in the activity label feature favoring recommendation \( p(A_2 = social|class = 1)/p(A_2 = social|class = 2) = 1.22 \) similar to ‘Movement’ \( (1.41) \) and ‘Leisure’ \( (1.43) \). Again, however, we also see strong individual preferences. For example, S50 does not favor Leisure \( (0.86) \), while S36 does not favor ‘Social’ activity \( (0.76) \) but instead activities labeled as ‘Food’ \( (4.43) \). The SVM model showed similar results for the beta weightings.

### 4.3 Simulation study

The offline analysis does not address the issue of introducing unknown (i.e., not previously encountered) activities, which is a key feature of a recommender system. Therefore, we explored this method of incorporating novelty into the recommendations through a simulation study. We labeled the 320 activities from PES [17] and used these as input to the personalized models. The goal was to test the personalized models by comparing the recommended activity with

<table>
<thead>
<tr>
<th>1% Holdout training size</th>
<th>10% Holdout training size</th>
<th>90% Holdout training size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>NB G</td>
</tr>
<tr>
<td>F1</td>
<td>0.59</td>
<td>0.56</td>
</tr>
<tr>
<td>Recall</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>Precision</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>AUC</td>
<td>0.50</td>
<td>0.58</td>
</tr>
<tr>
<td>Error</td>
<td>0.44</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Model results at 10, 30, and 80% holdout training size for the non-clinical sample. Base: Baseline model, P: Personalized model, G: General model

<table>
<thead>
<tr>
<th>10% Holdout training size</th>
<th>30% Holdout training size</th>
<th>80% Holdout training size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>NB G</td>
<td>NB P</td>
</tr>
<tr>
<td>F1</td>
<td>0.73</td>
<td>0.72</td>
</tr>
<tr>
<td>Recall</td>
<td>0.68</td>
<td>0.62</td>
</tr>
<tr>
<td>Precision</td>
<td>0.78</td>
<td>0.87</td>
</tr>
<tr>
<td>AUC</td>
<td>0.5</td>
<td>0.65</td>
</tr>
<tr>
<td>Error</td>
<td>0.34</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Fig. 8. Simulation Results: The top 5 recommendations from a list of 320 positive activities at 3, 10 and 90% holdout training size for P1 and P4

5 Discussion

We gathered high quality in-the-wild activity data to evaluate the feasibility of predicting positive activities. We followed best practice to utilize and evaluate textual information [37], [38], [43] resulting in the implementation of both an NB and an SVM recommender system. The models were highly consistent, and the information gain was similar between the features, allowing us to be confident about our conclusions.

Our main results show that the model is able to incrementally learn from past activity data to more accurately predict whether a given activity would be rated as positive by the user. The model also quickly converges to high accuracy suggesting this method could play an important part in supporting new digital approaches to mental health and well-being. Studies of activity and depression [17], [18] demonstrate that engaging in enjoyable activities has a positive health benefit improving mood and reducing depressive symptoms. In this study we were able to classify activities as positive while filtering out activities resulting in negative consequences. We have therefore proposed a recommender system with the potential to improve mental health and well-being.

Combining the results from the clinical and non-clinical sample we observe that the general model performs better than the personalized model when presented with few observations. However, this changes relatively quickly and with on average 58.92 (SD = 20.96) activities the personalized model outperforms the general with error rates at 30% on average with some participants reaching error rates of 10%. Our approach of recommending specific activities represents an improvement on the current state of the art. For example, Emotical [12] recommends to users that they schedule a type of activity (e.g. socialize, exercise), but our model has shown the feasibility of recommending specific activities ‘volunteering for a cause I support’. Given prior intervention work on the importance of generating specific recommendations for user compliance (e.g. [31]), this should improve the likelihood that participants will engage in the recommended activity. However, care needs to be taken in evaluating and deploying such models for this vulnerable target population. While the result suggests a relatively small number of activity entries for the models to converge, the requirement for data entry might be a barrier for some participants. One possible solution and an extension to the work might be to first profile new users using standard psychological surveys [12]. New participants could be matched to existing participants according to the similarity of their survey responses, and receive recommendations based on the models of closest matching participants.

Our modeling approach also offers transparency potentially supporting user insights from the trained recommender system. Providing users with visualizations such as that shown in Figure 5 might promote insights about the factors that influence users’ mood, allowing them to generate novel positive activities independently of our system. Such visualizations might also provide value for conversations between patient and therapist, offering insights for both, and allowing more productive planning of weekly activities. For instance, P5 seems to enjoy movement-related activities.
and these could be incorporated into their treatment schedule.

Finally, while our results are encouraging, given the vulnerability of potential user populations, it is vital to carefully stage the process of incorporating personalized recommendations directly into a mobile system. We demonstrated one initial approach to apply novel data through a simulation study (Section 4.3) using PES activities. The immediate findings look promising when compared with the trained features. This means that we are closer to predicting pleasurable activities according to the preferences of the user. This improves the accessibility of BA treatment. Although validated, however, such a probabilistic model does not address questions about the design of a working recommender application. We also do not advocate replacing therapist-guided activity planning with a purely system-driven approach. We acknowledge that there is a large gap between predicting positive activities and providing clinically useful treatment. We don’t expect to bridge that gap any time soon, but computational algorithms have the definite potential to improve the efficiency of clinical processes and generate new insights into mental health and well-being [44]. Our system might therefore provide support for weekly planning activities by suggesting personalized positive activities for discussion between patient and clinician. One first step, as emphasized by Sanches et al. when designing for mental health, is the use of HCI techniques [45]. Techniques that assist in exploring user needs, clinical constraints, understanding requirements and context, and the design space on e.g., user interactions. These would help us understand key areas such as whether the user might carry out the activity, or how a user perceives suggestions from a non-human entity.

5.1 Limitations and Future Work

We demonstrated the feasibility of recommending explicit activities. However, the analysis was done offline and in a simulated scenario. This is a natural consequence of the limited prior research in the area, and the high risk involved in offering experimental system-derived recommendations to vulnerable populations. A future step would be to validate the recommendations in a new in-the wild user study. In an “online” live context, we could obtain direct interactive feedback about the recommendations we generate. More specifically to evaluate the recommender model and the usefulness of its recommended activities, we intend to combine expert clinician’s evaluations with participant assessments. Such a study would also investigate which recommended activities the participant chose to carry out, and whether these activities are clinically beneficial. An initial approach to recommending activities might be those from PES, as demonstrated in the simulation study. However, in a recommender system, it is important to consider whether the recommended activities are actually carried out by the participant. Participants need to do more than scheduling recommended activities, they also need to act upon them. Hence, it is therefore also necessary to collect and incorporate follow-up data about whether activities were actually enacted.

In terms of the recommender model itself, a future step would be to design a context-aware model: Knowing the participant’s current location, the weather, or time of day might drastically change their positivity rating. “Going for a walk” very late at night, would not be the best idea. Instead “Snuggling up with a soft blanket” might be more suitable. We therefore plan to explore the feasibility of recommending activities taking into consideration the context of the user based on information collected from the participant’s phone, including location, time of day, and physical activity.

6 Conclusion

We have introduced a novel context for recommender systems, to recommend personal activities to improve personal mental health and well-being. We described how we collected high-quality data from two mobile applications and used this to model activity valence based on text and activity type features using two classification methods. In the context of the application, our results show promise given the prevalence of mental illness and the opportunity to contribute to novel forms of personalized treatment discreetly to anyone who has access to a smartphone. In addition, models were reasonably accurate, with personalized models outperforming general models, which argues for the utility of building user-specific models.

We believe the results of this work constitute an important step in developing new technologies that will help the vulnerable user group of depressed patients. We understand the potential ethical concerns that such recommender systems might bring. However, we claim that recommending positive activities can potentially offer valuable benefits to the behavioral therapy process.

Acknowledgments

This study was funded by the Innovation Fund Denmark through the RADMS project, Copenhagen Center for Health Technology (CACHET), and the Augustinus Foundation. The Emotical project was also partially supported by NSF Grant IIS 1321102. A special thanks to all the participants who used MORIBUS and EmotiCal day in, day out.

References

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2972007, IEEE Transactions on Emerging Topics in Computing

SPECIAL SECTION ON NEW FRONTIERS IN COMPUTING FOR NEXT-GENERATION HEALTHCARE SYSTEMS 10


Darius A. Rohani received his M.Sc. in Biomedical Engineering from the Technical University of Denmark (DTU). Currently he is finalizing his Ph.D. at the Department of Health Technology at DTU with Prof. Jakob E. Bardram as main supervisor. His research include mental health, behaviorism, cognitive restructuring, and user-centered design, with main technical expertise within supervised machine-learning, longitudinal analysis by Linear Mixed Effects Models, and statistical resampling methods.

Aaron Springer received his Doctorate in Computer Science from University of California Santa Cruz (UCSC) where he specialized in Human-Computer Interaction. His specific interests include human-centered AI and algorithmic bias.

Victoria Hollis received her Ph.D. in Cognitive Psychology from the University of California, Santa Cruz (UCSC). Her research interests include personal informatics, emotion self-perception, and authority effects of personal analytics.

Jakob E. Bardram is a professor in Computer Science at the Technical University of Denmark and is the director of the Copenhagen Center for Health Technology (www.cachet.dk). His research interests include software architecture for mobile and ubiquitous computing, mobile sensing, and Human-Computer Interaction. He holds a PhD degree from the University of Aarhus, Denmark.

Steve Whittaker is a professor in Human-Computer Interaction at the University of California Santa Cruz (UCSC), and Editor of the journal Human Computer Interaction. His research interests are designing human centric systems for personal informatics, including quantified self and personal information management. He holds a PhD from St. Andrews University, Scotland.