RESEARCH ARTICLE

Measuring happiness increases happiness

Jannik Roessler[1] · Peter A. Gloor[2]

Abstract

Happiness has been an overarching goal of mankind at least since Aristotle spoke of Eudaimonia. However measuring happiness has been elusive and until now has almost exclusively been done by asking survey questions about self-perceived happiness. We propose a novel approach, tracking happiness and stress through changes in body signals with a smartwatch, the "Happimeter". It predicts individual emotions from sensor data collected by an Android Wear smartwatch, such as acceleration, heartbeat, and activity. The Happimeter was used over three months in the innovation lab of a bank with 22 employees to measure individual happiness, activity, and stress. The participants were randomly divided into an experimental and a control group of similar size. Both groups wore the watch and entered their subjective happiness, activity and stress levels several times a day. The user-entered ratings were then used to train a machine learning system using the sensors of the smartwatch to subsequently automatically predict happiness, activity, and stress. The experimental group received ongoing feedback about their mood and which activity, sensor signals, or other people, made them happier or unhappier, while the control group did not get any feedback about their predicted and manually entered emotions. Just like in quantum physics we observed a "Heisenberg Effect", where the participants made aware of their measurements changed their behavior: Members of the experimental group that received happiness feedback were 16% happier, and 26% more active than the control group at the end of the experiment. No effect was observed for stress.

Keywords Emotion prediction · Wearable technology · Smartwatch · Machine learning · Heisenberg effect

✉ Jannik Roessler
 roessler@wim.uni-koeln.de

 Peter A. Gloor
 pgloor@mit.edu

1 Cologne Institute for Information Systems, University of Cologne, Cologne, Germany

2 Center for Collective Intelligence, Massachusetts Institute of Technology, Boston, USA

 ⊘ Springer

Introduction

Most people think they know when they are happy, sad, or stressed. However, emotional awareness is far from given [1]. People are notoriously bad of recognizing when they are sad or stressed. Being aware of one's own emotions is of great help in coping with the challenges of daily life. Fear gets the driver on the highway to slam on the brakes when the truck is getting too close, but it can also be a powerful inhibitor when there is no reason to be fearful [2]. On the positive emotional side, striving for happiness is one of the biggest and most important goals of human beings [3]. In general, happiness increases productivity [4], work performance [5], career success [6] as well as health and other social factors [6]. On the other hand, unhappiness hampers efficiency because employees start acting sloppily [4]. Further, a negative mood can lower expectancy, instrumentality, and valence for rewards and thus lead to a worse work performance [7].

But how do we know when we are happy? Previous studies mainly surveyed and interviewed individuals in order to access their overall well-being [5, 6, 8]. However, all of these approaches suffer from problems of self-assessment and cognitive bias [9]. Further, such methods are not only time-consuming and expensive, but they also lack the ability of real-time analysis and actions.

A system is required which automatically tracks the mood of a person at any time of the day to circumvent these limitations of surveys and interviews. Thanks to the rise of wearable sensor technologies such as smartwatches and wristbands, we get access to the most important source of emotional information: the body. Not only do smartwatches count the steps and give immediate feedback to the wearer on the display, but their built-in sensors access body and context information like heart rate, noise, or acceleration. Previous studies have shown that machine learning can be used to extract high-level features from body sensors that achieve high accuracy in different areas like activity recognition [10] or emotion recognition via an "Emotive Couch" [11]. Further, Blanchflower and Oswald [12] showed that physiological information i.e. blood pressure can be used to assess well-being. Thus, in this paper we introduce an alternative way to measure emotions through machine learning using body sensors from a smartwatch.

To make individuals aware of their emotions and increase their well-being we employ a process called "virtual mirroring". It has been introduced in [13], referring to a process in which individuals are shown metrics of their own communication behavior while they are told which communication behavior is desirable. In earlier work doing virtual mirroring by analyzing e-mail communication, it has been shown that individuals will change their behavior to be more collaborative, productive, and innovative [14]. Thus, through virtual mirroring organizations can enhance their performance and outcome.

Heisenberg's uncertainty principle states that two complementing characteristics, such as location and movement, cannot be determined simultaneously. Trying to measure a definite position will make the measurement of the movement less accurate, and vice versa [15]. Gloor [16] has concluded that the same applies to humans. When a group of people is observed and if they know this,

their self-reflection is triggered which leads to a change in behavior. Furthermore, the measurement is not correct anymore due to people's change in behavior.

Based on these findings, it is therefore desirable to (a) measure and predict happiness, stress, and activity to make individuals aware of their positive or negative feelings, and (b) find ways to increase individual well-being.

Related work

Being happy is fundamental for most humans, and in most societies, happiness is a highly valued goal [17]. "In all periods in history, and in all countries, a large majority of people expresses their desire to live a happy life" [3]. However, it is hard to describe the term happiness because there exists a wide variety of definitions [18]. Robertson and Cooper [6] describe happiness in terms of physical, social and psychological (mental) well-being. Further, the authors allege that the latter consists of two components: First, experiencing positive emotions and a feeling of happiness. Second, experiencing purpose and meaning in what we are doing with our lives. Frey and Alois [3] claim that well-being "refers to the satisfaction people gain when they reach desired goals". Moreover, the authors provide different determinants of happiness e.g. income, autonomy, creativity, personal relationships, health, to be involved in decision making processes and socio-demographic factors. From an organizational research perspective, happiness-related constructs are spread over several levels, from transient level to personal level to unit level [19]. At the transient level, individuals experience happiness as a result of an instantaneous action or an event e.g. the enjoyment of a task or the momentary mood at work. The personal level characterizes individuals' happiness in general such that individuals can be differentiated from each other e.g. typical mood at work or job satisfaction. The unit level describes happiness of groups of individuals such as a team or the organization as a whole e.g. covering collective job satisfaction or the group mood. Further, Fisher [19] alleges that all these levels not only vary in stability over time but also depend on their specific context.

Most studies use self-report questionnaires to measure happiness [6]. One of the key instruments is a survey of subjective life-satisfaction which assesses an individual's overall subjective well-being by asking questions such as "Taken overall, how satisfied are you with the life you lead on a scale from 0 (totally dissatisfied) to 10 (totally satisfied)?" [3]. Another method is the U-Index, which measures the time during the day when the individual was in an unpleasant state. In contrast to the survey of subjective life satisfaction, this approach has no personal scale because one's individual intensity of unpleasantness is not taken into account [9]. Other methods, such as the Day Reconstruction Method (DRM) [20] or the Experience Sampling Method [3] try to measure emotional aspects in moment-to-moment situations. The latter asks individuals randomly and multiple times a day about their well-being while in DRM each participant needs to list their activities of the previous day, including information about the time, place and field of activity. To determine the emotions during each activity, participants have to select the intensity of predefined emotions on an ordinal scale.

Another measurement is the Positive and Negative Affect Schedule [21]. It contains two 10-item scales to measure positive and negative effects. Participants are then asked to provide a ranking for each item on a five-point scale of 1 (very slightly) to 5 (extremely). Several studies have attempted to elucidate the complex relationship between happiness and work performance using the aforementioned approaches. The results are various ramifications of happiness, especially for businesses.

For example, Amabile et al. [5] concluded that being happy at work is associated with more creativity and proactivity. The authors conducted a multi-study longitudinal research program with 222 students. The participation ranged from 9 to 38 weeks due to individual projects length. Each individual had to answer a demographic questionnaire in the beginning as well as one questionnaire per day. The authors employed the Electronic Event Sampling Methodology [22] which is an adaptation of the Experience Sampling Methodology [3].

Further, in [4] 713 individuals were exposed to four different experiments, each investigating if happiness is increasing productivity. In their experiments the authors used comedy movie clips, chocolate, fruit and drinks as well as talking about tragedies in order to make people more or less happy. As expected, the results show that happiness is indeed increasing productivity. Other studies confirm these findings e.g. [5, 20]. Amabile and Kramer [23] found that "People do better work when they are happy".

Robertson and Cooper [6] showed that higher levels of well-being are not just linked to higher levels of income, better work performance and higher career success, but also to higher levels of more successful marriages and friendships and better health. The authors used several case studies to underpin their findings, describing them all would go beyond the scope of this work.

According to Baron et al. [24], happy employees have less interpersonal conflicts. This finding was confirmed by two studies. In the first study, 87 students were either provoked or not during a business case. Afterwards the participants were exposed to one of five treatments: control, gift, self-deprecation, flattery or excessive flattery. In a last step, individuals completed a questionnaire about reactions and feelings to the negotiation as well as about their happiness. The results show that conflicts may sometimes be reduced through happiness triggered by positive incentives such as flattery, gift or self-deprecating remarks by an opponent. However, the authors conducted another study to further support the argument. Again, the task was to negotiate a business case but this time, each participant was treated with either humor, flattery or nothing (control). Each treatment was presented three times: before, during and after the negotiation. In a last step, the students completed a questionnaire similar to the one of the first study.

Although all of these cases assessed happiness, they suffer from problems of self-assessment and cognitive bias [9]. Further, such questionnaire-based methods are not only time-consuming and expensive, but they also lack the ability of real-time analysis and actions. Often, questionnaire results are enriched by taking into account information obtained from interviews with individuals [6], which makes the analysis even more time-consuming and expensive. Finally, the results can be unreliable if employees fear that the boss will read the questionnaire [25].

A few studies exist which used the abundance of information collected by wearable devices, but they were never tested in a real world, working environment. For example, Khan and Lawo [26] collected various physiological information by using a pulse sensor and an eHealth platform, including the following sensors: 2D accelerometer, blood pressure, oxygen in the blood, body temperature, airflow, electrocardiogram (ECG), electromyography and GSR. While wearing these devices, the 24 participants were shown 100 images one after another. These images were divided into four emotional subgroups, namely: Joy, dislike, sad and stress. After looking at five images of the same subgroup, the participants had to choose one emotion from the aforementioned emotional states, extended by the terms 'normal' and 'no idea'. Subsequently, the authors created models for recognizing different emotional states by considering the provided answers from the participants and body sensor information.

Nguyen et al. [27] built an application for emotion prediction by using a Mio Link Heart Rate Wristband, which was connected to a smartphone by Bluetooth Low Energy (BLE). The device collected the heart rate continuously in the background, while the five participants recorded their emotions via the application whenever possible. Nguyen et al. considered six different emotional states including: Fear, anger, sadness, disgust, neutral and happiness. Further, they distinguished between three different subgroups: negative emotions (fear, anger, disgust, and sadness), neutral emotions (neutral) and positive emotions (happiness). They created different feature sets by calculating maximum, mean, median as well as discrete wavelet transformation and Mel-frequency cepstral coefficients (MFCC) from the heart rate.

The most similar work to ours is [25]. The authors used various sensors i.e. accelerometer, location, sound-level, infra-red signals, and temperature combined in a badge which was worn on the chest. The badges were worn by 468 office workers from 10 organizations for 5000 days. They used the Center for Epidemiologic Studies Depression Scale which asks about typical symptoms of depression to assess happiness. Participants had to answer 20 questions by using a four-point scale. The authors discovered a correlation between physical motion and collective happiness. However, the work is different from ours in three points. First, happiness was measured on an organizational level rather than on an individual level. Second, happiness was measured weekly rather than multiple times a day. Third, instead of using a complex questionnaire, we implemented an Experience-based Sampling Method [3] with only three questions directly asked to the user on the watch.

Research methodology

The happimeter

The Happimeter [28] is a system for tracking and predicting human's mood. It provides feedback to individuals about their mood and what influences it. Our hypothesis to be tested in this experiment is if this feedback will make them actually happier. The Happimeter uses a smartwatch for collecting the necessary information and a website, as well as a phone application for providing detailed insights.
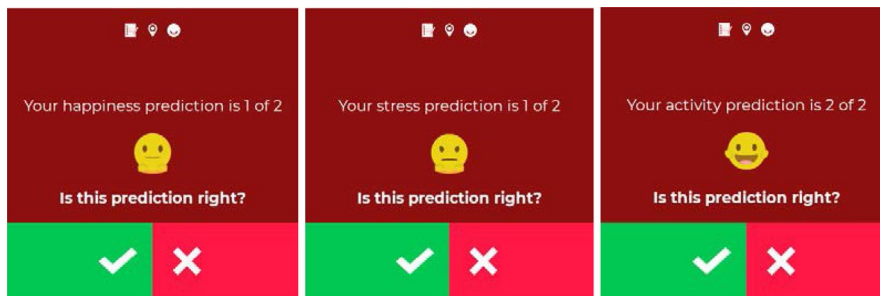
Fig. 1 The three different questions asked by the watch. Users can either confirm or deny and correct the suggested mood

The smartwatch is the most important component of the system as it is responsible for collecting the body sensing data. For the experiment described in this paper, users were equipped with either a Mobvoi TicWatch S2, a Mobvoi TicWatch E2 or a Mobvoi Ticwatch S. These watches were running on WearOS. All three kinds of watches had the following sensors. First, an accelerometer sensor which measured the acceleration force in m/s^2 on three physical axes, namely: x, y, and z. Thus, when the user moved the watch, e.g. shaking or tilting it, an acceleration to one of the given axes was caused. Second, a step counter which recorded the number of steps since the last sensor measurement. Third, a heart rate sensor which collected the heart signals in beats per minute. Fourth, a microphone which recorded the noise expressed by the amplitude; and finally, a position sensor which determined the user's location by localizing the longitude, latitude, and altitude. Each sensor was running automatically for 30 s every 20 min trying to collect as many measurements as possible. Further, exogenous variables i.e. weather data and time-related data were added by the system based on the GPS information provided by the watch.

In order to build a customized individual model for each user, not only the physiological information is necessary but also subjective information about the user's mood to train the system. Thus, a survey is integrated into the Happimeter application which asks users approximately every two hours about their activity level, happiness and stress. Simultaneously, this survey is predicting the mood level, providing the wearers with information about what the system thinks how they feel. Users can either confirm or reject and correct the suggested mood. For example, users are asked: "Your happiness prediction is 1 of 2. Is this prediction correct?" If not, they can enter the correct value on a sliding scale from zero to two, where zero is representing that the user is not happy, one that the user is happy, and two that the user is very happy. The same applies for activity and stress (see Fig. 1 for an illustration).

The predictions are based on a machine learning algorithm which uses the user-entered mood and smartwarch sensor data. While new users start with a generic model for the prediction, individualized models are trained over time based on the feedback of the user. The difference between a generic model and an individualized model is that the latter is based on the user's mood input and the former uses the mood input of every user in the system. Note that users can
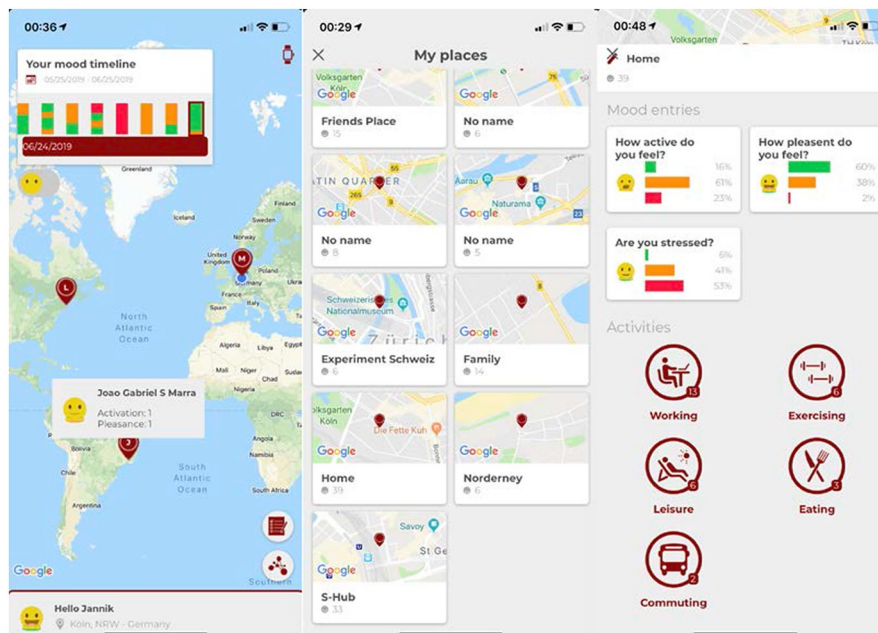
Fig. 2 Illustration of the Happimeter smartphone app. From left to right. **a** Users have an overview over their last survey inputs. For example, on the 24th of June the user has been mostly happy. Further, they can see the location and mood of their friends. The big letters are representing the user's friends. **b** Users can name places where they usually make inputs. **c** For each place they receive detailed information such as the average of their happiness, activity, and stress

enter their sensor and mood data manually in addition to the predictions generated automatically, further increasing the accuracy of the system.

A second key component of the Happimeter system for conducting virtual mirroring is the smartphone application which is available for Android and Apple. Although users can also enter their mood data there, the primary purpose of this component is to monitor emotions and to manage the social network. Users can look at their past survey inputs examining how happy they have been on a specific day (see Fig. 2a). Further, users can evaluate their mood input associated with the geolocation. They can label specific places like "Home" or "Friends Place". Such a function allows users to accurately analyze how they felt in certain places (see Fig. 2b, c).

Furthermore, the app can be used for managing the social network including functions such as accepting friend requests, sending friend requests, controlling privacy options for mood sharing and unfriend somebody. Users have not only access to a friend's happiness, activity, and stress level but can also be given access to the geolocation (see Fig. 2a).

Finally, the application is sending notifications to the user if she or he is not happy according to the Happimeter predictions. Based on the system's prediction of happiness, activity, and stress, different notifications are sent. For example, if

the system predicts that the user is very unhappy, very active and very stressed, it suggests taking a long nap. The default interventions are shown in Fig. 10.

The last component of the Happimeter system is the website. It can be used for reviewing the collected measures and mood inputs. On the dashboard users can see their mood inputs over the last 30 days (see Fig. 11), as well as the system's prediction based on the user's last sensor data. Another function is to provide insights about the driver of a user's mood. Users can see what variables influence their mood the most (see Fig. 12). Further, they can see by whom they are influenced and on whom they exert influence (see Fig. 13). Moreover, the website can be utilized for modifying the default interventions such that each user can incorporate individual notifications. Finally, a user can create a team consisting of multiple users. For each team the mood inputs are aggregated, the average is calculated, and the information is displayed on the website.

Experimental setup

A real-world experiment was conducted with 22 full-time employees of the Sparkassen Innovation Hub (S-Hub) in Hamburg, Germany from May 1, 2019 to August 19, 2019. Due to allergic reactions with the strap, four employees aborted the experiment prematurely after one month leading to 18 remaining participants, of whom 16 (88.89%) were male and 2 (11.11%) female. Every participant was equipped with a smartwatch at the beginning of the experiment. After setting up the watches each participant was given a few days to get used to the system. Furthermore, this time was necessary to replace the generic happiness prediction models with individualized ones.

To verify the reliability of happiness, activity, and stress prediction, one model was created for each question for each user. Sensor data of each participant was collected approximately every twenty minutes during working hours. Note that sometimes the participants forgot to wear their watches or the sensors failed to record any information because of hardware or software issues. Further, the watch was asking participants approximately every two hours about their happiness, activity, and stress. Both sensor and mood data were used to create models by using different machine learning algorithms i.e. Decision tree, gradient boosting with decision trees, random forest, support vector machine (SVM), feedforward neural network and long short-term memory (LSTM) models. Due to the availability of user-entered information about the user's mood, the algorithms could easily be evaluated by comparing real mood values and predicted mood values.

To explore if feedback and recommendations affect happiness, stress and activity—i.e. the Heisenberg effect through virtual mirroring indeed exists—the participants were equally divided into two groups, a control group and an experimental group. The latter had access to all functionalities mentioned above. Members of the experimental group received recommendations, based on their predictions. These recommendations were sent approximately every three hours during business hours. Further, they had insights about their drivers, past mood inputs and geolocations, as well as they could use the social network functionalities. In contrast, the control
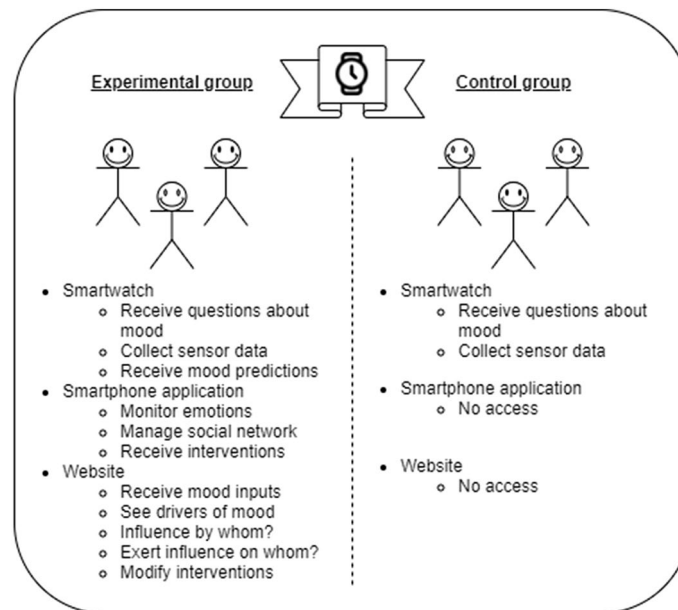
Fig. 3 Comparison of control and experimental group. While the experimental group had access to all of the Happimeter's feature like receiving mood predictions, receiving interventions and monitoring emotions, members of the control group did not have access to any functionalities except for receiving questions about mood and collecting sensor data

group had none of the Happimeter functionalities. Users were just asked to wear the watches to provide sensor data and to answer the questions to provide mood data. Their access to the website was denied. Functionalities in the application on the phone were restricted such that participants could not see their last mood inputs. Further, when asking about the mood the watch did not provide any predictions. Instead members of the control group were only asked about their happiness, activity and stress levels (see Fig. 3 for an illustration).

Feature selection

Feature selection is crucial in machine learning, as it can have a significant impact on the outcome of the algorithms. In this paper recursive feature elimination (RFE) [29] was used to create a feature set. In RFE, the classifier is first trained with all features. Next, the least important feature is removed. That procedure is repeated until a single feature is left. The best feature set is the one that achieves the highest accuracy. For the creation of the final feature set, a decision tree was used as an estimator. This algorithm can return the feature importance of each feature, which, in turn, can be used for determining the best features.

RFE was applied on all available features including the nine features which resulted from the five sensors of the Mobvoi smartwatches described above, plus

Table 1 Best feature set using RFE

Method	Sensor
Minimum, maximum	Heart rate, microphone
Average	Accelerometer X, accelerometer Y, accelerometer Z, heartrate, microphone
Magnitude	Number of steps, temperature, humidity, pressure, hour, weekday, session

weather information, i.e. temperature, humidity, pressure, wind, clouds, and weather, as well as time-related information, i.e. hour, weekday and session. The latter determined the part of the day. The best feature set using RFE based on a decision tree is presented in Table 1.

One model was built for each question for each user, i.e., each user had three different models: One for predicting happiness, one for predicting activity, and one for predicting stress. In order to build these models, it was necessary to apply some preprocessing steps. These steps are explained in the following.

To create the training and testing instances, each mood data entry was combined with one instance of sensor data, by searching for the closest timestamp in the sensor data that matched the mood data. Unfortunately, a match could not always be found between mood and sensor data, and thus, some information could not be considered while fitting the models. The answers for the different questions were used as labels.

During the preprocessing step, those columns, which either had more 'not-a-number' (nan) values than numerical values or where the standard deviation of the column was zero, were dropped. Remaining nan values in columns which had not been dropped were replaced with the mean of the respective column. Further, one weather variable had to be encoded such that the values were numerical. To cope with the imbalanced dataset, *cost-sensitive learning* [29] was taken into account when building the models. Additionally, *tenfold cross-validation* with *stratification* was used when evaluating the generalization performance of each algorithm.

Performance measures

Different performance metrics, i.e. *accuracy*, *F1-score* and a *combination of both* were calculated based on the different feature-sets, periods and questions to evaluate and compare each algorithm. Although accuracy is a common evaluation metric, in this case it might not be a good indicator because when the class distribution in a dataset is skewed, the accuracy can be very high which leads to the assumption that the model is doing well [30]. The F1-score, a combination of precision[1] and recall[2]

---

[1] $\frac{\text{true positives}}{\text{true positives}+\text{false positives}}$ [29].

[2] $\frac{\text{true positives}}{\text{true positives}+\text{false negatives}}$ [29].

was used for evaluating the model to overcome the drawback of accuracy. Mathematically, it is expressed as:

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

Finally, to compare different models, a single number metric was necessary, which takes into account accuracy and F1-score equally. Without such a metric, one model could be superior w.r.t accuracy while another could be superior w.r.t F1-score. The combination was simply the mean of both metrics:

$$\frac{\text{Accuracy} + \text{F1}}{2} \tag{2}$$

The metrics were calculated for all three models of each user. However, for the final evaluation, the various performances of each user were summarized, and the average was taken.

It was necessary to build a baseline against which the models' performance could be compared to infer the meaning of these metrics. Such a baseline can evince performance improvements while pointing out superior algorithms. Due to the highly imbalanced dataset it was reasonable to use a majority classifier, also called zeroR, as a baseline classifier. It is a naïve classifier that always predicts the most frequent class in the training set [30]. In the bottom line all the learned models should at least outperform the majority classifier.

Data analysis and results

Evaluate prediction

In total, 10,830 sensor data items were recorded during the experiment. 4845 sensor data items were collected by the experimental group, and 5985 sensor data items were collected by the control group.

In contrast to the sensor data, the mood data was manually entered by the user. As already described above, users were asked about their happiness, stress, and activity. The answers along with the question id and other information such as timestamp, longitude, and latitude were sent to the server. In total 6844 mood data entries were recorded, 2123 happiness answers, 2126 activity answers and 2115 stress answers. 4034 mood data records were collected by the experimental group, and 2810 mood data records were collected by the control group. Note that the answers, given by the users, were imbalanced such that most of the users stated that they had been active, happy and stressed. The answers' ratios for activity, happiness, and stress were 341:1191:594, 192:1325:606, 811:1122:182, respectively.

Table 2 shows the key performances of each algorithm for predicting activity. The random forest was most accurate. Its combined value of 76.12% was 14,25% higher than the baseline performance as mentioned in section 2.4 and 1.72% higher than the combined value of the second-best algorithm, namely,

Table 2 Performance metrics of each algorithm for predicting activity

	Activity		
	Accuracy (%)	F1-score (%)	Combination (%)
Baseline	67.75	55.99	61.87
Decision tree	68.54	66.75	67.65
Gradient boosting	75.78	73.02	74.40
Random forest	**79.06**	**73.18**	**76.12**
SVM	77.42	69.79	73.60
FFNN	68.07	68.07	68.07
LSTM	69.57	68.97	69.27

The best metrics are written in bold

Table 3 Performance metrics of each algorithm for predicting happiness

	Happiness		
	Accuracy (%)	F1-score (%)	Combination (%)
Baseline	82.87	76.04	79.46
Decision tree	83.54	82.85	83.20
Gradient boosting	87.42	85.13	86.27
Random forest	**88.42**	**86.18**	**87.30**
SVM	87.42	82.37	84.89
FFNN	64.74	64.74	64.74
LSTM	49.45	47.91	48.68

The best metrics are written in bold

gradient boosting. While the F1-score of random forest and gradient boosting was almost the same, they differed in the achieved accuracy. Here, the latter was outperformed by 3.31%. Decision tree, FFNN, and LSTM performed worst, and they were mere slightly above the baseline performance.

Table 3 lists the best performance metrics of each algorithm for predicting happiness. Again, random forest performed the best. Its combined value with 87.30% was superior to the combined value of the baseline majority classifier by 7.84%. Although the accuracy of the baseline performance was already very high, the random forest outperformed it with 5.55%. Gradient boosting was barely inferior of random forest. The deep learning algorithms performed worst. Both algorithms achieved a combined value even worse than the combined value of the baseline performance.

Table 4 lists the performance metrics of each algorithm for predicting stress. Here, SVM was the best algorithm. This classifier was exceeding the baseline performance by 9.97% w.r.t the combined value. Interestingly, SVM's F1-score was worse than most other algorithms. However, its accuracy was outperforming the accuracy of the other algorithms and thus, in sum its performance was slightly superior. However, gradient boosting, random forest, decision tree, and

Table 4 Performance metrics of each algorithm for predicting stress

	Stress		
	Accuracy (%)	F1-score (%)	Combination (%)
Baseline	65.10	52.39	58.74
Decision Tree	67.41	66.51	66.96
Gradient Boosting	70.45	66.85	68.65
Random Forest	70.09	66.46	68.28
SVM	**73.94**	**63.47**	**68.71**
FFNN	66.11	66.11	66.11
LSTM	54.19	52.09	53.14

The best metrics are written in bold

Table 5 Happiness comparison between experimental ($N=9$) and control group ($N=9$)

	Average happiness	Shapiro–Wilk (p-value)	Mann Whitney U (p-value)	Number of observations	Standard deviation	Minimum	Maximum
Self-reported							
Experimental group	1.26	1.17e−41	5.30e−09	1218	0.54	0	2
Control group	1.10	2.53e-33		905	0.62	0	2
Predicted							
Experimental group	1.29	0	0	89,684	0.57	0	2
Control group	1.13	0		87,084	0.62	0	2

FFNN just performed slightly worse than the SVM. The second worst algorithm—ignoring the baseline—was the FFNN which achieved a combined value of 66,11%. The difference in the combined value makes clear how close the algorithms were to each other: The FFNN was excelled by merely 2.6%.

Evaluating Heisenberg's effect

To investigate whether receiving feedback about what drives one's dissatisfaction, and suggestions about how to improve it will increase employee happiness, the average happiness of experimental and control group was calculated. The assumption was that feedback is making the user more conscious of his/her feelings while the recommendations should actively help the user to become happier and more satisfied. Consequently, happiness should be higher in the experimental group than in the control group. Although some of the participants had worn the watches outside of working hours, only business days from 4 am to 10 pm were included in this analysis, leaving 6364 mood data entries.
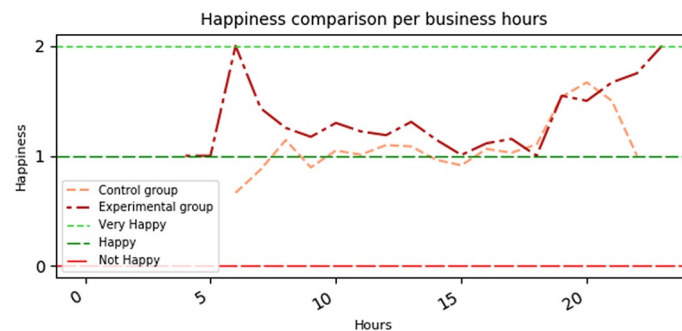
Fig. 4 Self-reported happiness per business hour

Happiness

Table 5 resumes the differences in happiness between experimental and control group, with happiness values both manually entered, and predicted by the machine learning system. We find that, indeed as hypothesized, the experimental group, which had been getting detailed happiness feedback on watch, phone app, and website, shows 15% higher happiness. This means the virtual mirroring seems to work, and getting happiness feedback will indeed increase happiness.

Figure 4 showing the average happiness per business hour, illustrates this finding. The experimental group was always happier than the control group with one exception: At 8 pm the control group was slightly happier than the experimental group. Note that this illustration looks more skewed because, at very early or late business hours, the number of employees making a mood input was significantly smaller than during the day. This made the happiness average more dependent on outliers among individuals rather than on the entire group.

Figure 5 illustrates happiness averaged over days and weeks, supporting the hypothesis too. Both graphics underline that on average the experimental group was happier than the control group. Further, two other characteristics are conspicuous. First, the control group experienced a happiness low on August 2nd, August 6th, and August 8th (see also the week from 5 to 9 August). Second, in the last week of the observation, the average happiness in both groups increased rapidly.

Activity

A similar picture emerges for activity (Table 6). Again, average activity over the total observation period in the experimental group was 0.26 higher than in the control group (1.23 and 0.97, respectively).

Figure 6 illustrates how the activity changed during the business hours for both groups. The experimental group was always more active than the control group. Both groups experienced an activity low at the end of the day.

Figure 7 compares the average activity of both groups in different temporal perspectives. Again, the results support the hypothesis since, in both graphics, the
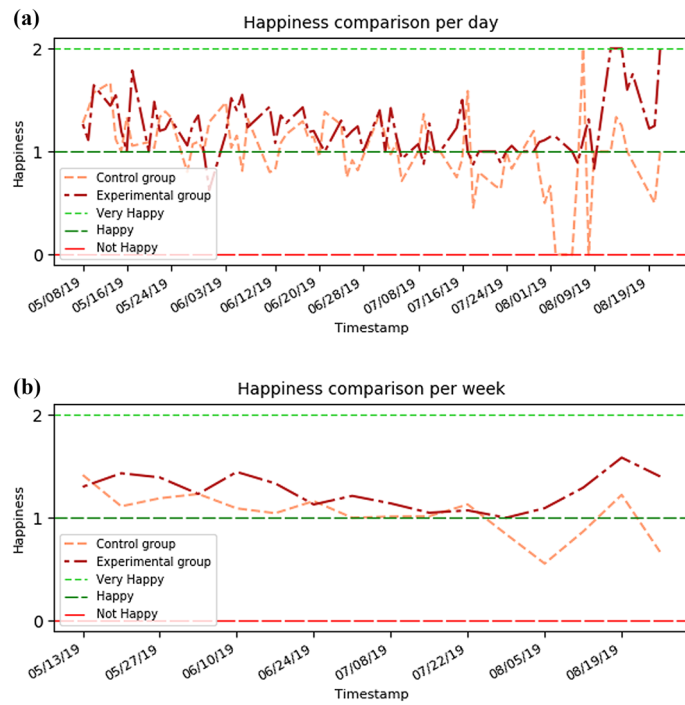
(a)

(b)

Fig. 5 Comparison of average self-reported happiness over time. **a** Average happiness per day. **b** Average happiness per week

Table 6 Activity comparison between experimental ($N=9$) and control group ($N=9$)

	Average activity	Shapiro–Wilk (p-value)	Mann Whitney U (p-value)	Number of observations	Standard deviation	Minimum	Maximum
Self-reported							
Experimental group	1.23	9.05e−40	2.05e−18	1219	0.58	0	2
Control group	0.97	1.87e−31		907	0.71	0	2
Predicted							
Experimental group	1.14	0	0	89,681	0.49	0	2
Control group	0.95	0		87,085	0.63	0	2

experimental group was more active than the control group. Only between July 8th and July 17th the control group was more active. Further, at the end of the experiment, the experimental group was much more active on average than the control group.
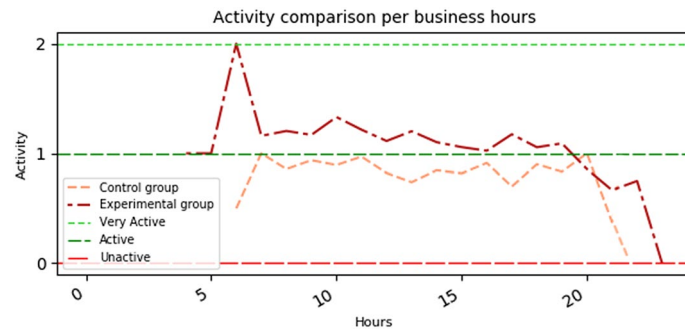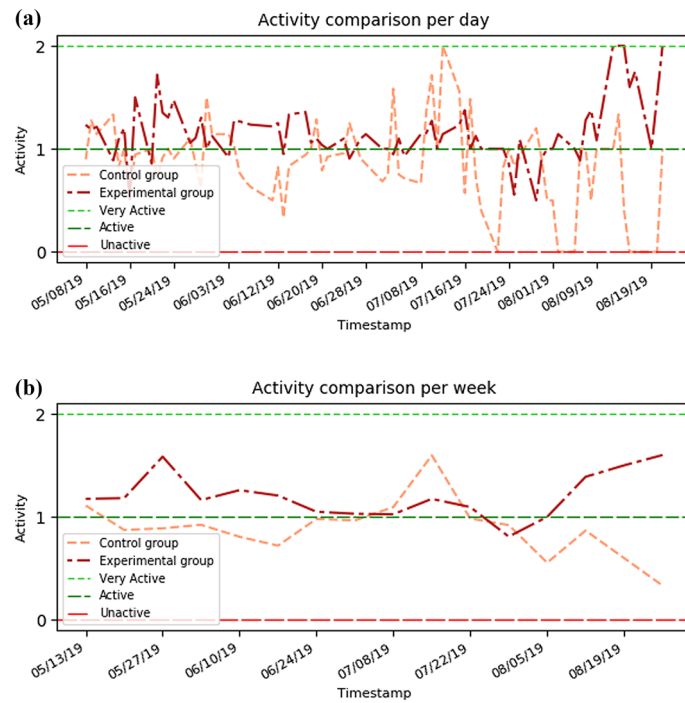
Fig. 6 Self-reported activity per business hour

Fig. 7 Comparison in average self-reported activity over time. **a** Average activity per day. **b** Average activity per week

Stress

Finally, the stress average over the total observation time was calculated for both groups. While the experimental group obtained an average stress value of 0.69, the control group achieved an average stress value of 0.72. Although users of the control group experienced more stress than the users of the experimental group, the difference was not significant (Table 7).

Table 7 Stress comparison between experimental ($N=9$) and control group ($N=9$)

	Average stress	Shapiro–Wilk (p-value)	Mann whitney U (p-value)	Number of observations	Standard deviation	Minimum	Maximum
Self-reported							
Experimental group	0.69	7.62e−40	0.66	1213	0.58	0	2
Control group	0.72	3.11e−33		902	0.66	0	2
Predicted							
Experimental group	0.67	0	0	89,779	0.60	0	2
Control group	0.75	0		87,173	0.60	0	2
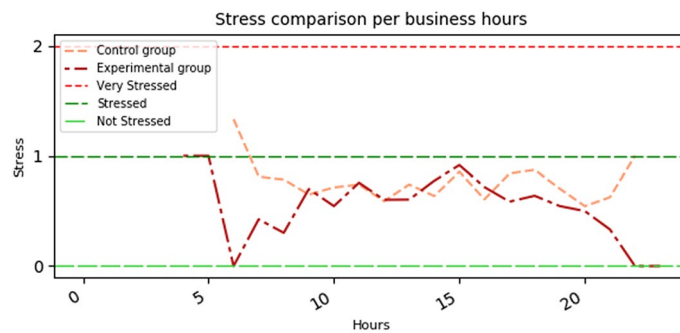
Fig. 8 Average stress during business hours

As Fig. 8 illustrates, the stress levels of the control group and the experimental group were running mostly in parallel during business hours. Again, in this illustration, the differences between the experimental and control group look more extreme. However, not all users reported stress values at every business hour. Thus, the stress average could have been dependent on individual persons rather than on the entire group.

Figure 9a and b show more clearly that the stress curves of both groups were approximately the same. Figure 9a shows the change of stress over the days of the entire observation period. It can be seen that sometimes, the stress in the experimental group was higher than in the control group and vice versa. Figure 9b shows the change of stress over the weeks of the entire observation period. This graph clearly shows how little the two groups differed from each other.
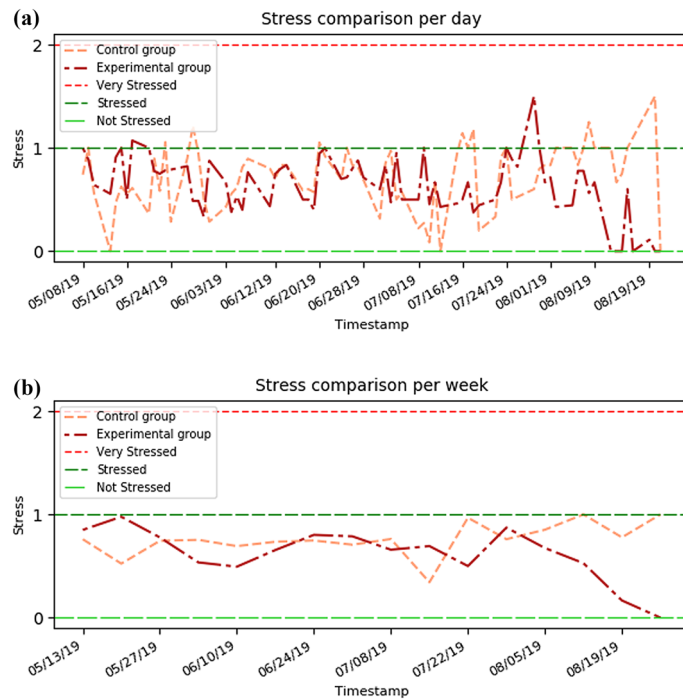
(a)

(b)

Fig. 9 Comparison in average self-reported stress over time. **a** Average stress per day. **b** Average stress per week

Discussion

In this research, a smartwatch-based system was introduced, which used physiological information, such as heart rate or ambient noise, exogenous variables such as weather, and machine learning algorithms, to reliably predict people's satisfaction in terms of happiness, activity, and stress. Moreover, it was shown, that predictions, made by the system, can be used to provide feedback and recommendation, which, in turn, can increase satisfaction by increasing happiness and activity.

The results of this work demonstrate practical implications of such a system for managers. It was shown that sensor-based systems reliable predict happiness, activity, and stress. This information can be used to make employees more satisfied with their job due to different factors. First, managers can use our approach to make employees happier and more active either by mirroring back to people what makes them happy and active like we did in the experiment or by observing the employees and recommending interventions for increasing personal satisfaction. The same approach can also be used to make the employees less stressed. Ultimately, such a working environment, were everybody is aware of what makes them happy, active and less stressed, will increase productivity and foster organizational benefits. Additionally, companies can get rid of surveys and time-consuming interviews by using an automated system like the one described in this paper which can predict employee

satisfaction automatically. This not only saves money but also enables real-time satisfaction tracking. Nevertheless, companies need to be aware of the sensibility of the data. They not only have to ensure data security and privacy, but they also must be careful not to misuse the information in any, for the employees, harmful way.

It seems that physiological information as well as weather- and time-related information can be used to assess human's mood. Further, Heisenberg's uncertainty principle and virtual mirroring might be used to change the behavior of individuals by triggering their self-reflection. Note that the participants were neither forced to read the feedback nor to actively put the interventions into practice. No mechanism was implemented to control how the users perceived the recommendations. Participants in our experiment were enthusiastic about our system, but the question remains how other, more conservative companies might be encouraged to use similar systems.

There are some weaknesses of the current study. First, it might be that the participants of the control group were less happy and less active than the participants of the experimental group only because they did not receive any feedback and recommendation. It might be that the participants of the control group felt inferior because of such an experimental set up, which led to more unhappiness and inactivity. However, we assume that this would overrate the influence of our Happimeter system, as the work in the office should be the focus and key influence on employee satisfaction. Second, the number of participants was limited. More attendees would have reinforced the significance of the result. Further, the number of male participants was higher than the number of female participants. Thus, other experiments, with better gender balance, are needed in order to further verifiy the relationship between physiological information and activity, happiness and stress across all genders. Finally, the system did not operate perfectly. Some information got lost because of hardware malfunction or software issues.

In future work, the Happimeter system should be implemented in other companies and with more participants. The system should be further developed to reliably predict happiness, activity, and stress. Maybe, the system could also be put in another context such as using physiological information to predict whether somebody is going to buy a car or not. Also, it would be interesting to see how another smartwatch and thus, other sensors, could be used in combination with the system.

Appendix

See Figs. 10, 11, 12, 13.

Id	Pleasance	Activation	Stress	Suggestion
1	Very unhappy	Very unactive	Stressed	Take a 10 minutes walk.
2	Very unhappy	Unactive	Stressed	Meditate for 10 minutes.
3	Unhappy	Very active	Stressed	Lie down for 5 minutes.
4	Very unhappy	Very unactive	Relaxed	Whistle a tune.
5	Very unhappy	Very unactive	Very stressed	Write down your problems.
6	Very unhappy	Very active	Relaxed	Drink a cup of tea.
7	Very unhappy	Unactive	Very stressed	Take ten deep breaths.
8	Very unhappy	Very active	Very stressed	Take a long nap.
9	Unhappy	Very active	Very stressed	Listen to a song.
10	Unhappy	Very unactive	Stressed	Take a short walk.
11	Unhappy	Very unactive	Very stressed	Discuss your problems with a friend.
12	Unhappy	Unactive	Stressed	Meditate for 5 minutes.
13	Unhappy	Very active	Relaxed	Drink a glass of water.
14	Unhappy	Unactive	Very stressed	Take a deep breath.
15	Unhappy	Very active	Very stressed	Take a short nap.
16	Very unhappy	Unactive	Relaxed	Make somebody else happy.
17	Unhappy	Unactive	Relaxed	Give a small gift.

Fig. 10 Default suggestions for increasing happiness based on low stress, activation, happiness levels

Fig. 11 Overview of last mood inputs on the website. The white lines represent the individual values manually entered by the user, and the grey line represents the average of all users in the system
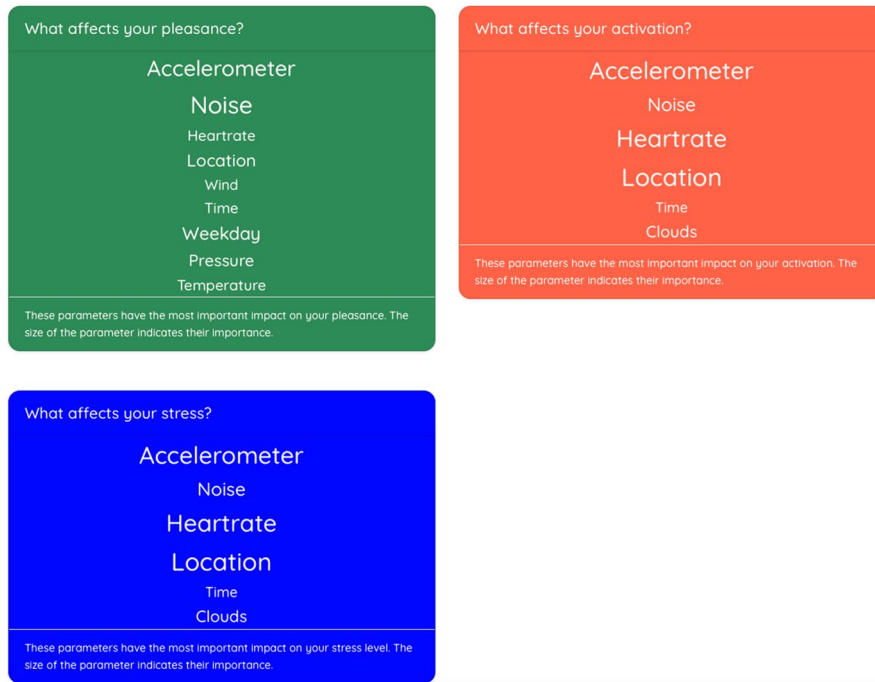
Fig. 12 Website showing drivers of activity, happiness, and stress

Fig. 13 Website shwing who is influencing the mood of a person

References

1. Ekman, P. (Ed.). (2008). *Emotional awareness: Overcoming the obstacles to psychological balance and compassion: A conversation between the Dalai Lama and Paul Ekman*. New York: Times Books.
2. Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford: Oxford University Press.
3. Frey, B. S., & Stutzer, A. (2018). *Economics of happiness*. New York: Springer International Publishing.
4. Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and Productivity. *Journal of Labor Economics, 33*(4), 789–822.
5. Amabile, T. M., Barsade, S. G., Mueller, J. S., & Staw, B. M. (2005). Affect and creativity at work. *Administrative Science Quarterly, 50*(3), 367–403.
6. Robertson, I., & Cooper, C. (2011). *Well-being: Productivity and happiness at work*. London: Palgrave Macmillan.
7. Judge, T. A., & Kammeyer-Mueller, J. D. (2011). Happiness as a societal value. *Academy of Management Perspectives, 25*(1), 30–41.
8. Sharifzadeh, M., & Almaraz, J. (2014). Happiness and productivity in the workplace. *American Journal of Management, 14*(4), 19–26.
9. Kahneman, D., & Krueger, A. B. (2006). Developments in the measurement of subjective well-being. *Journal of Economic Perspectives, 20*(1), 3–24.
10. Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters, 119*, 3–11.
11. Rus, S., Joshi, D., Braun, A., & Kuijper, A. (2018). The emotive couch: Learning emotions by capacitively sensed. *Procedia Computer Science, 130*, 263–270.

12. Blanchflower, D. G., & Oswald, A. J. (2011). International happiness: A new view on the measure of performance. *Academy of Management Perspectives, 25*(1), 6–22.

13. Gloor, P. A. (2017). *Sociometrics and human relationships: Analyzing Social networks to manage brands, predict trends, and improve organizational performance* (1st ed.). Bingley: Emerald Publishing.

14. Gloor, P., Colladon, A. F., Giacomelli, G., Saran, T., & Grippa, F. (2017). The impact of virtual mirroring on customer satisfaction. *Journal of Business Research, 75*, 67–76.

15. Heisenberg, W. (1927). Über den anschaulichen: Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik, 43*(3–4), 172–198.

16. Gloor, P. A. (2017). *Swarm leadership and the collective mind: Using Collaborative innovation networks to build a better business*. Bingley: Emerald Publishing.

17. Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American psychologist, 55*(1), 34.

18. Kesebir, P., & Diener, E. (2009). In pursuit of happiness: Empirical answers to philosophical questions. In *The science of well-being* (pp. 59–74). Dordrecht: Springer.

19. Fisher, C. D. (2010). Happiness at work. *International Journal of Management Reviews, 12*(4), 384–412.

20. Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science, 306*(5702), 1776–1780.

21. Watson, D., Clark, L. A., & Carey, G. (1988). Positive and negative affectivity and their relation to anxiety and depressive disorders. *Journal of Abnormal Psychology, 97*(3), 346.

22. Amabile, T. M., Whitney, D., Weinstock, J., Miller, L., & Fallang, C. (1997). *"What Really Happens in Creative Projects: Event Sampling Through Electronic Data Collection."* Harvard Business School Working Paper, No. 98-036.

23. Amabile, T., & Kramer, S. (2011). *The progress principle: Using small wins to ignite joy, engagement, and creativity at work*. Boston: Harvard Business Review Press.

24. Baron, R. A., Fortin, S. P., Frei, R. L., Hauver, L. A., & Shack, M. L. (1990). Reducing organizational conflict: The role of socially-induced positive affect. *International Journal of Conflict Management., 1*, 133–152.

25. Yano, K., Akitomi, T., Ara, K., Watanabe, J., Tsuji, S., Sato, N., … Moriwaki, N. (2015, June). Profiting from IoT: The key is very-large-scale happiness integration. In *2015 Symposium on VLSI Technology (VLSI Technology)* (pp. C24-C27). IEEE.

26. Khan, A. M., & Lawo, M. (2016, September). Developing a system for recognizing the emotional states using physiological devices. In *2016 12th International Conference on Intelligent Environments (IE)* (pp. 48–53). IEEE.

27. Nguyen, N. T., Nguyen, N. V., Tran, M. H. T., & Nguyen, B. T. (2017, October). A potential approach for emotion prediction using heart rate signals. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 221–226). IEEE.

28. Gloor, P. A., Colladon, A. F., Grippa, F., Budner, P., & Eirich, J. (2018). Aristotle said "happiness is a state of activity"—Predicting mood through body sensing with smartwatches. *Journal of Systems Science and Systems Engineering, 27*(5), 586–612.

29. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Amsterdam: Elsevier/Morgan Kaufmann.

30. Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking* (1st ed.). Sebastopol, CA: O'Reilly Media.