Inferring emotional valence using wearable fitness tracker and psychometrics tests

[Green, Mikella; Menendez, Joaquin & Zhao, Sicong]

Background

As the popularity and capability of wearable activity trackers has increased over the last 20 years, so has an interest in what their derived data can tell us about our overall health. Commercially available wearable fitness trackers use accelerometers and heart rate monitors to give users insight regarding sleep, daily activity, and heart rate which are all important variables for a healthy life. Research has shown that our daily emotions also play an important role in our overall health, with large cumulative amounts of negative emotions being detrimental to physical, psychological, and behavioral health [1]. Given this, finding a way for people to monitor their emotional valence, or how positive or negative their mood is,

An accurate emotion detection system would benefit multiple parties. For wearable fitness tracker manufacturers, emotion detection capability could make their products more competitive in the global market. The global emotion detection and recognition market was valued at $12.37 billion in 2018 and is expected to reach a value of $91.67 billion by 2024, at a CAGR of 40.46% [2]. Industries including consumer economics and marketing indicate a growing demand from both the commercial and consumer market. [3] From a customer's perspective, an increased awareness of how daily stimuli affect their mood could be beneficial to their mental health. Considering the potential benefits, our project explores the feasibility of combining emotion detection with wearable fitness trackers. **The aim of our project is two-fold: first, we examine if we can recognize emotional valence, using data derived from commercially available wearable fitness trackers and second, we seek to see if the inclusion of additional indicators including psychometric tests and demographic data improves the performance of our model.**

Previous Research

Physiological signals of emotion. Emotional arousal involves changes in the autonomic system which gives rise to changes in multiple physiological signals, including heart rate, respiration, and perspiration [4]. Psychophysiology studies have used non-invasive biosensors to map specific patterns of physiological signals to emotional states. Multiple studies suggest that there is some degree of specificity of autonomic nervous system arousal to different emotions, but that effects are context-specific and can be affected by different factors including induction paradigms, personality, and demographic characteristics [5]. Moreover, valence-specific patterning was found to be more consistent than emotion-specific patterning: negative emotions were associated with stronger autonomic responses than positive emotions [6].

Emotion detection systems. Several studies have demonstrated the feasibility of stress detection from wearable biosensor data [7,8,9]. These previous stress detection systems have used physiological signals like heart rate variability, galvanic skin conductance, and respiration rate. One obvious limitation of previous work is that the biosensors used were designed for research purposes and are not practical. They are obstructive to daily activities, not easily wearable, and easily confounded by naturalistic environments.

More recent studies have attempted to address the shortcomings of previous research, for example Bogomolov [10] and colleagues built a daily stress recognition system from mobile phone data, weather conditions, and individual traits. While the study didn't utilize fitness tracker data, it did demonstrate the feasibility of emotion detection outside of the laboratory environment. A 2018 study by Lawanont [11] and colleagues used wearable activity trackers and smartphones to build a daily stress recognition system based on physical activity and heart rate data. Their recognition model used heart rate, resting heart rate, number of steps, calories burned, and sleep quality. Our project aims to address the limitations of current work and also build upon current knowledge.

Data

Data Collection

The data for our project was collected by researchers affiliated with Dr. Gregory Samanez-Larkin through two studies conducted at Vanderbilt University in Nashville, TN and Yale University in New Haven, CT. Although the data originated from two different studies, only the measures that were common across the two studies were included. Participants completed a battery of self-report surveys that included demographic questions, psychometric surveys, and cognitive screenings. Next, participants completed a physical examination. Participants were given a Fitbit Charge HR device to be worn for the next 10 days. Over the 10 day period, participants' mood ratings were collected at three random times each day through an experience sampling survey link sent via SMS. For a more detailed description of the data see Table 5 in the appendix.

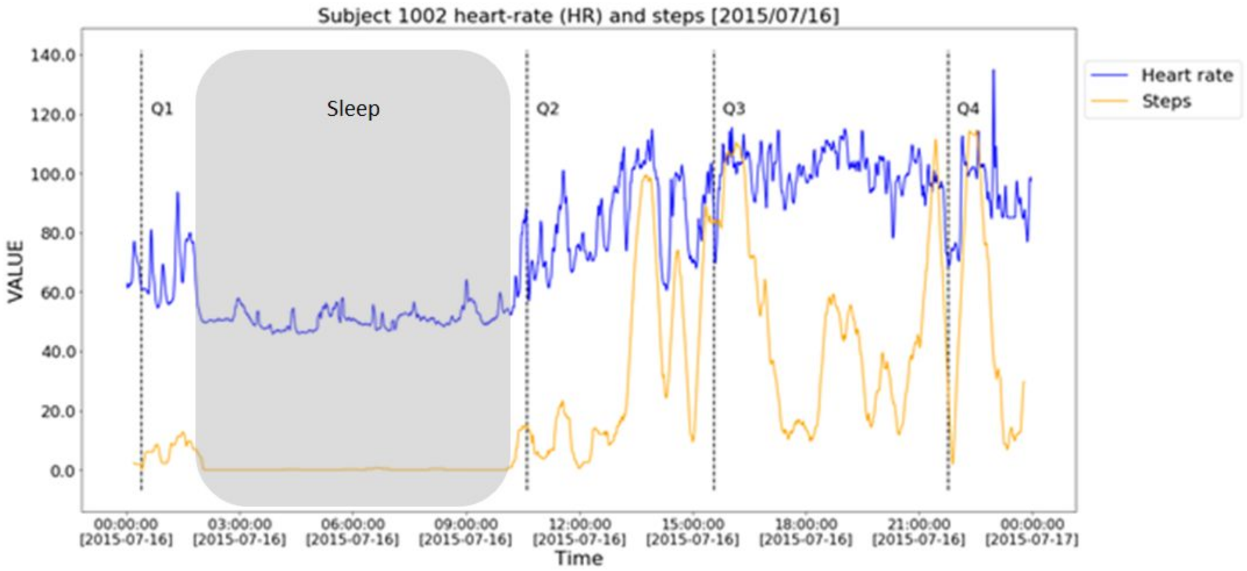


Fig 1. The figure shows a register of a single day of one of the subjects. The dashed line (Qn) shows when the subject completed an 'experiencing sample survey'. Steps (orange line) are measured on steps per minute, Heart rate (blue line) is measured on heart rate per minute.

Target Variables

We used the data from the psychometrics surveys and demographic data of every subject with their respective Heart and Steps data to infer his/her emotional states. According to the Circumplex Emotion model [6] emotion could be divided into two dimensions: Arousal and Valence.

Arousal could be defined as the activation level of the autonomic system. On the other hand, Valence is the affective quality referring to the intrinsic attractiveness (positive valence) or averseness (negative valence) of an event, object, or situation. As we mentioned before, valence-specific patterning tends to be more consistent than arousal for differentiate emotions, this is the reason why we decided to try to predict only the emotional valence of the subjects. In other words, we decided to focus if a subject was experiencing a pleasant or aversive emotional state rather than a high or low arousal estate.

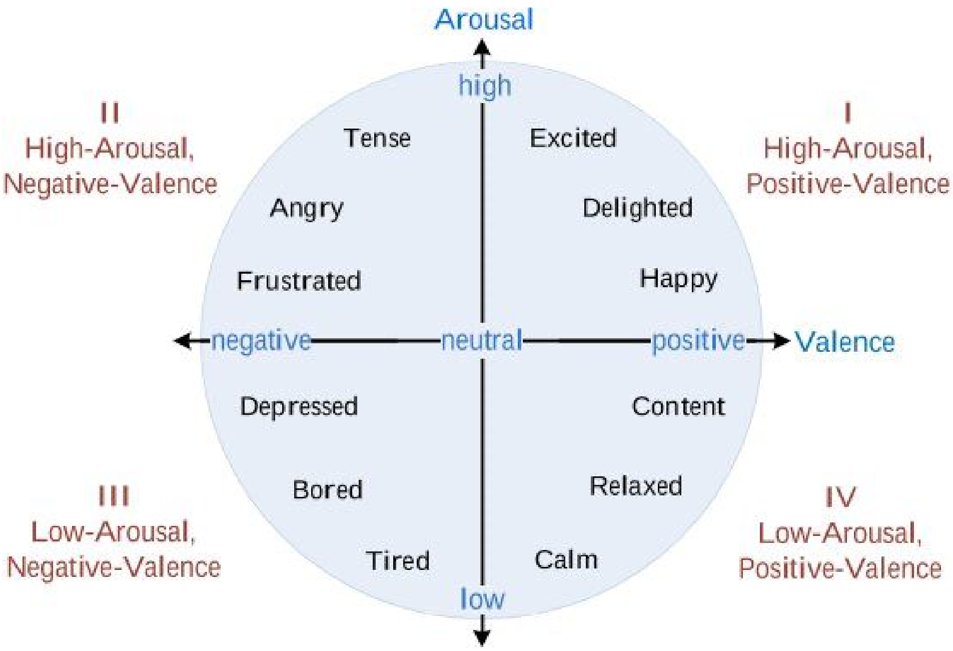


Fig 2. Core emotions established in the circumplex model. Extracted from [12]

In order to measure the emotional valence we used the subjects' answers to the Experiencing Sample survey. This survey consisted of 8 questions with a Likert format (value between 1 to 5) asking about how the subject was feeling at that moment (i.e. *'How calm , at rest, or serene do you feel right now?*'). Every question asked for a specific combination of valence and arousal (see Fig 2). Two questions measured only Valence: 'Positive' and 'Negative' (abbreviated as 'p' and 'n'), two questions measured only Arousal: 'Low Arousal' and 'High Arousal' (abbreviated as 'ha' and 'la') and four questions asked about a different emotion-arousal combination: 'Low Arousal Positive','High Arousal Positive', 'High Arousal Negative','Low Arousal Negative' (abbreviated as 'la_p', 'ha_p', 'ha_n', 'la_n').

We used this information to try to answer questions. First, can we know if a subject is experiencing a positive or negative emotional state?. To be able to answer this question we used the variables Low Arousal Positive,High Arousal Positive, High Arousal Negative, Low Arousal Negative, Negative and Positive (see Table 1), to construct a metric called 'Positive/Negative Emotion'.

Type of data	Variable name	Description	Measure Type	Sampling rate	Mean (SD)
Experience sampling survey	la_p	Low Arousal Positive (Likert scale)	Self report score	3 per Day (avg)	2.99 (1.11)
	ha_p	High Arousal Positive (Likert scale)	Self report score	3 per Day (avg)	2.34 (1.09)

	ha_n	High Arousal Negative (Likert scale)	Self report score	3 per Day (avg)	1.33 (0.7)
	la_n	Low Arousal Negative (Likert scale)	Self report score	3 per Day (avg)	1.91 (1.04)
	la	Low Arousal Likert scale)	Self report score	3 per Day (avg)	2.34 (1.09)
	ha	High Arousal (Likert scale)	Self report score	3 per Day (avg)	1.38 (0.74)
	n	Negative (Likert scale)	Self report score	3 per Day (avg)	1.35 (0.72)
	p	Positive (Likert scale)	Self report score	3 per Day (avg)	3.16 (1.08)
	start_survey	Timestamp when the subjects started answering the survey	Timestamp	3 per Day (avg)	-
	survey_no	# of surveys completed by the subject so far.	# of surveys per subject	3 per Day (avg)	22.5 (6.5)
	Period_of_day	Morning', 'Afternoon', 'Evening'	Category	3 per Day (avg)	-

Table 1. Variables from the Experience Sampling Survey. The variables 'la_p', 'ha_p', 'ha_n', 'la_n', 'p', 'n' were used to compose the two different metrics we tried to predict.

Positive Negative Emotion is a binary measure that labels the data into ‘Negative’ (1) or ‘Positive’ (0) depending on the answers reported by the subject. In other words, if the biggest valence value of any ‘negative’ variable is bigger or equal than the biggest ‘positive’ variable then we claim that the subject experienced a negative emotional state. This metric was calculated for every survey for all subjects.(See Formula 1)

$$\text{Positive Negative Emotion} = \begin{cases} 1, & \max(la\_n_{s,i}, n_{s,i}, ha\_n_{s,i}) \geq \max(la\_p_{s,i}, p_{s,i}, ha\_p_{s,i}) \\ 0, & \text{otherwise} \end{cases}$$

Formula 1. Where ‘*s*’ represents the subject and *‘i’* represents the survey number.

Our second question was: Can we predict if the subject is going to be happier or unhappier than his or her baseline levels? To answer this question we constructed another measure called

‘Normalized valence’. As for the ‘Positive Negative Emotion’ metric we only used the variables in the Experience sample survey that were measuring some type of valence. Normalized valence was calculated in two steps. The first step was to calculate a ‘Valence’ score for every survey. In this case Valence is an absolute value of how positive or negative is the subject’s emotional state. Then we used this value to calculate Normalized valence. The normalized valence metric compares the Valence score of every survey for every subject and compares it with his/her valence score baseline level. In other words, Normalized valence is an individual metric that informs if the subject reported an emotional state more positive or negative than his/her baseline level (See Formula 2).

$$\text{Valence} = \frac{\text{High Arousal Positive}_{s,i} + \text{Low Arousal Positive}_{s,i} + \text{Positive}_{s,i}}{3} - \frac{\text{High Arousal Negative}_{s,i} + \text{Low Arousal Negative}_{s,i} + \text{Negative}_{s,i}}{3}$$

$$\text{Normalized valence} = \begin{cases} 1, & Valence_{s,i} - Median(Valence_s) < 0) \\ 0, & \text{otherwise} \end{cases}$$

Formula 2. Where *‘s’* represents the subject and *‘i’* represents the survey number

Variable name	Description	Measure Type	Sampling rate
Target Variable	Valence	General metric of how positive or negative is the subject’s emotional state	Float
	Normalized valence per subject	Individual metric that informs if the subject reported an emotional state more positive or negative than his/her baseline level	Float
	Positive/Negative Emotion	Binary measure that labels the reported data in the experience sampling survey into ‘Negative’ (1) or ‘Positive’ (0).	Category

Table 2. Metrics constructed using the data from the Experience Sampling Survey. For this project we tried to predict ‘Normalized valence per subject’ and ‘Positive/Negative Emotion’ in new and current users.

Data Summary

After the data processing the total number of subjects was 83 with an average of 22.5 surveys answered per subject (SD = +- 6.5). The final dataset was composed of a total of 1873 observations with 180 features.

EDA

Distribution of subjects' responses to the Experiencing Sample survey

The first step on our EDA was to observe the distribution of the responses to the Experiencing Sample survey. As we mentioned before we only employed 6 variables from this survey (‘la_p', 'ha_p', 'ha_n', 'la_n', 'p' and 'n'). To clarify, positive score (p_score) is the mean of an individuals’ ratings for high arousal positive (ha_p), positive (p), and low arousal positive (la_p). Similarly to p_score, negative score (n_score) is the average of high arousal negative (ha_n), negative (n), and low arousal negative (la_n) ratings. Both p_ and n_score indicate the valence score of a participant.

As figure 3 indicates, the ‘ha_n’ and `n` metrics are very right skewed, this means most of the subjects were reporting they had low scores in negative emotion. This could be the results of positivity bias, which means people have a tendency to report positive emotion rather than negative emotion. More specifically,

- About 80% subjects reported `ha_n` equal to 1
- About 76% subjects reported `n` equal to 1
- About 68% subjects reported `ha_n` and `n` both equal to 1

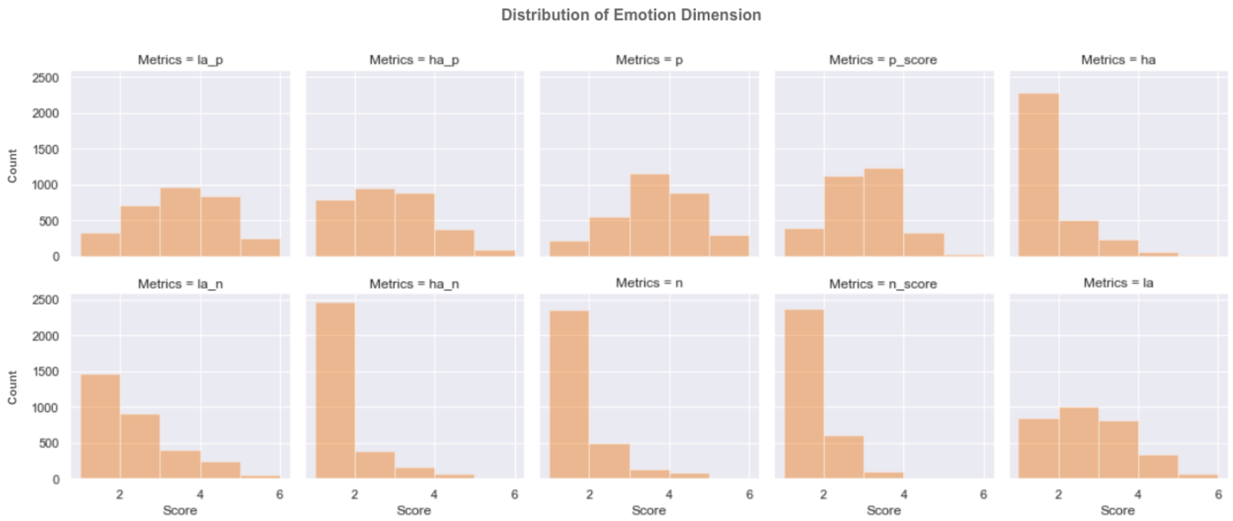


Fig 3. Distribution of answers for the different variables of the Experiencing Sample survey. The variables ‘p_score’ and ‘n_score’ were calculated by mean(ha_p, p, la_p) and mean(ha_n, n, la_n) respectively.

There are multiple ways to use the emotional state labels to construct the binary outcome that indicates if the subject is happy or unhappy. The first approach is comparing the mean value of positive measure (ha_p, p, la_p) and negative measure (ha_n, n, la_n) to decide if a subject is happy or unhappy. The second approach is comparing the maximum value of positive measure (ha_p, p, la_p) and negative measure (ha_n, n, la_n) to decide if a subject is happy or unhappy. The interpretation can vary across these two labels. The first label represents the average emotion is positive or negative, while the second label represents the strongest emotion is positive or negative. To us, we are more interested in predicting the strongest emotion of our user. Because this would give us a higher chance to identify users in desperate situations, who truly need instant help.

Additionally, the distribution for the labels from the second approach (26.6% observation with label ‘unhappy’’) is more balanced than the labels from the first approach (10.8% observation with label ‘unhappy’’). This makes our further analysis using the second approach to generate the labels easier than using the first approach to generate the labels.

By Subject Analysis

The subject level statistics indicate that there is a large amount of variance across subjects. Below is the distribution of the individual mean and standard deviation of positive valence score and negative valence score. As we can see in figure 4, these statistics are widely spread when compared with the value range of ‘p_score’ and ‘n_score’ (from 1 to 5)

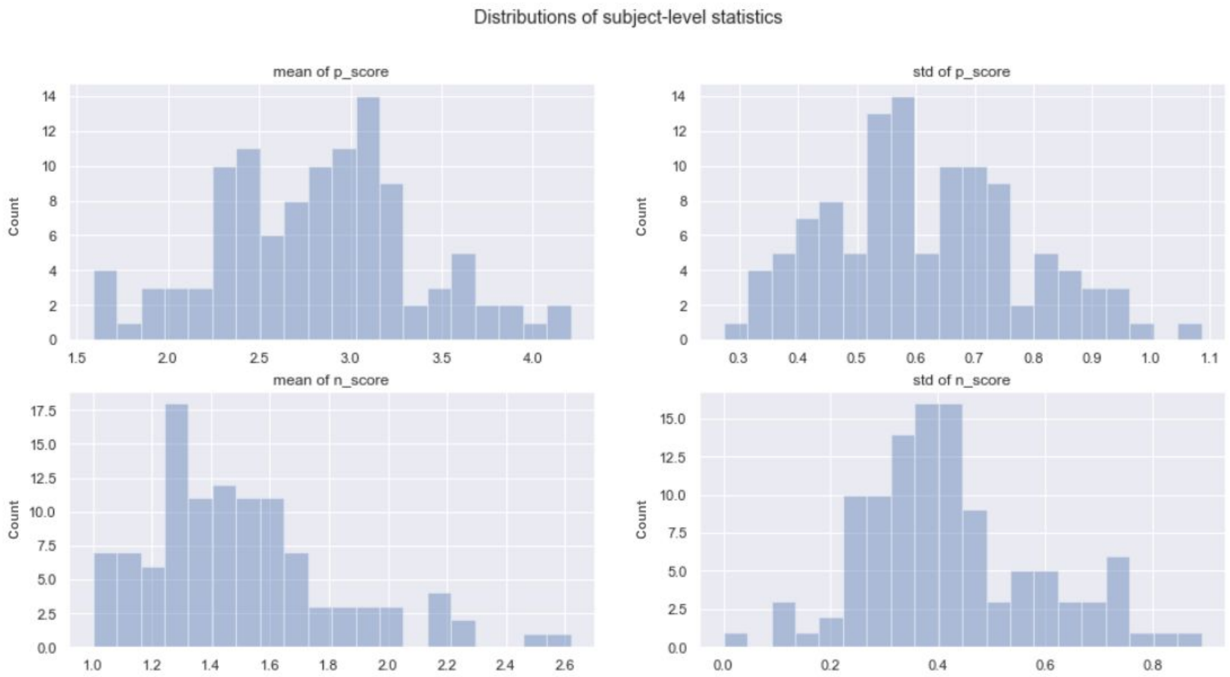


Fig 4. Distribution of mean and standard deviation for ‘p_score’ and ‘n_score’

As a result of this analysis, we have decided to normalize the valence by subjects, then build another type of models to predict if a subject is happier or unhappier than his/her normal mood (see Formula 2).

Model & Important Findings

We have 1872 observations with 173 features, which is not necessarily a large dataset. Therefore, even though emotional states can be complicated, flexible models like neural networks are not a good approach for this specific problem. Given the amount of data we have, tree based methods and simple models are logical choices.

For tree based models, we used CatBoost, a gradient boosting on decision trees library with categorical features support. And Random Forest, a tree-based ensemble learning method. For the simple model, we used Logistic Regression as a comparison.

To figure out which model is the most suitable one, we modeled for positive/negative emotion with current users using all the data, with Linear Regression, Random Forest and CatBoost. And it turns out CatBoost reported the highest F1 score between the 3 models tested. Therefore, we decided to move on with CatBoost.

Additionally, an imbalanced dataset can harm the model performance because of the severely skewed class distribution and the unequal misclassification costs. To address this problem, we have integrated four methods. These sampling approaches will be applied to our positive/negative valence binary label, which has 26.6% of the observations with label ‘unhappy’ and 73.4% observations with label ‘happy’.

1) Synthetic Minority Oversampling Technique (SMOTE), a statistical technique for increasing the number of cases by generating new instances from existing minority cases.
2) Up sample the minority cases.
3) Down sample the majority cases.
4) Assign class weight to loss function, in this way to balance the influence of two classes to our models.

These sampling approaches will be applied to our positive/negative valence binary label, which has 26.6% of the observations with label ‘unhappy’ and 73.4% observations with label ‘happy’.

In terms of the goals of this study, as illustrated in the Data section, we have 2 target variables representing two research goals. The target variable is ‘Positive Negative Emotion’, which is decided by the strongest emotion among all 6 metrics (la_p, p, ha_p, la_n, n, ha_n) to be positive or negative. The other target variable is ‘Normalized valence per subject’. By looking into this target variable, we are predicting whether a subject is happier or unhappier than his/her normal mood.

Application wise, after discussing with our partner, we recognized 2 scenarios of interest and two type of models under each scenario, which includes

1) Predict x for current users with only Fitbit data, where x could be positive/negative emotion or normalized valence binary label.
2) Predict x for new users with only Fitbit data, where x could be positive/negative emotion or normalized valence binary label.
3) Predict x for current users using all data, where x could be positive/negative emotion or normalized valence binary label.
4) Predict x for new users using all data, where x could be positive/negative emotion or normalized valence binary label.

For each target variable, we have looked into these 4 scenarios.

Model Type 1: Prediction for Positive/Negative Emotion

Table 3 shows the performance evaluation of the best models for predicting **Positive Negative Emotion.** As a comparison, the precision for random guessing is 0.273.

	Scenario	**Data**	**Sampling Methods**	**Recall**	**Precision**	**F1**

1	Current Users	Fitbit	SMOTE	0.550	0.347	0.425
2	New Users	Fitbit	Assign Class Weight to Loss Function	0.419	0.228	0.295
3	Current Users	All Data	Assign Class Weight to Loss Function	0.626	0.482	0.544
4	New Users	All Data	SMOTE	0.870	0.260	0.399

Table 3. Summary of best models for predicting ‘Positive Negative Emotion’

As shown in table 3, the model that gives the highest F1 score is the one predicting positive/negative emotion for current users using all data with class weight assignment. It has achieved an F1 score to be 0.544, and a precision to be 0.482, improved by 76.6% than random guesses. In order to understand each feature's influence on the prediction, we investigated the Shapley value of each feature in this model (see figure 5). The goal of Shapley value is to explain the prediction of an instance x by computing the contribution of each feature to the prediction.

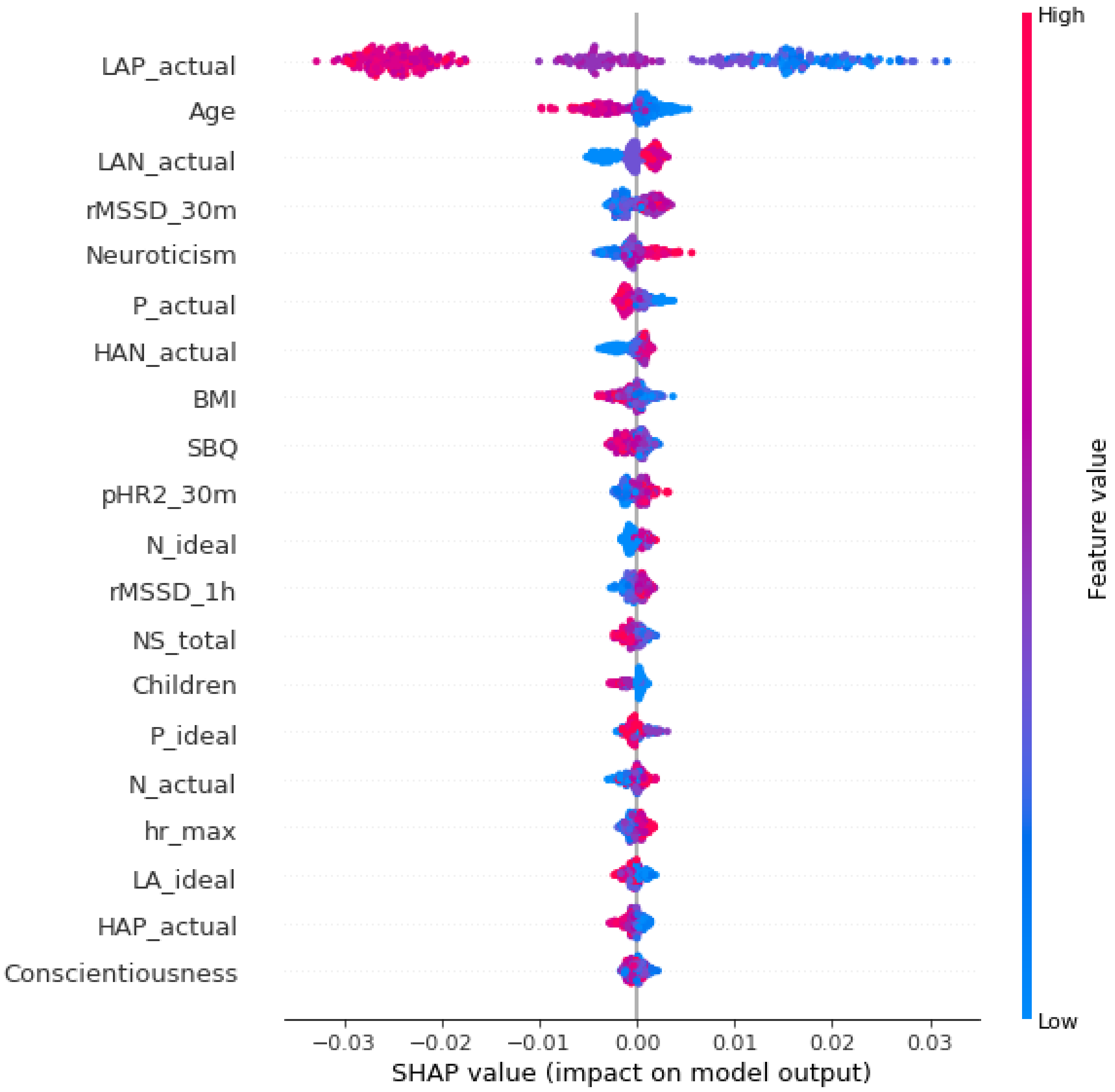


Fig 5. The Shapley Values of the features from the best model that predict positive/negative emotion**.** There are 3 things which help you understand figure 6: (1) Each row represents a feature. (2) The color indicates relative value for each feature. blue represents relative lower value, red represents relative higher value. (3) The value on the axis at the bottom represents the influence on the output value, which is the probability that a subject is unhappy. Besides, the values LAP_actual, LAN_actual, P_actual, HAP_actual, LA_ideal, P_ideal, HAN_actual, N_actual, LAP_idea correspond to variables from the Affect Valuation Index test (see Table 5 at the appendix).

Through analysis, there are three findings from table 3 and figure 5.

1) Using additional data (psychological, demographic data, physical health data) helps improve the prediction. In both scenarios, models trained with all data give better performance. F1 score improved by 0.119 and 0.104 for current users and new users respectively.
2) The precision of prediction for new users is lower than the precision of random guesses.
3) Psychological data are very important. The 4 most important features are:

a) Previous emotional states (LAP_actual, LAN_actual, P_actual, HAP_actual)
b) Age
c) Heart rate feature rMSSD_30m
d) Neuroticism

Model Type 2: Prediction for Relative Valence

Table 4 shows the performance evaluation of the best models for predicting **'Normalized valence per subject'.** As a comparison, the precision for random guessing is 0.564.

	Scenario	Data	Sampling Methods	Recall	Precision	F1
1	Current Users	Fitbit	-	0.578	0.617	0.596
2	New Users	Fitbit	-	0.699	0.558	0.617
3	Current Users	All Data	-	0.877	0.590	0.705
4	New Users	All Data	-	0.959	0.555	0.703

Table 4. Summary of best models for predicting 'Normalized valence per subject'

As shown in table 4, the model that gives the highest F1 score is the one predicting relative valence for current users using all data with class weight assignment. It has achieved an F1 score to be 0.705, and a precision to be 0.590, improved by 4.6% than random guesses. In order to understand each feature's influence on the prediction, we investigated the Shapley value of each feature in this model (see figure 7).

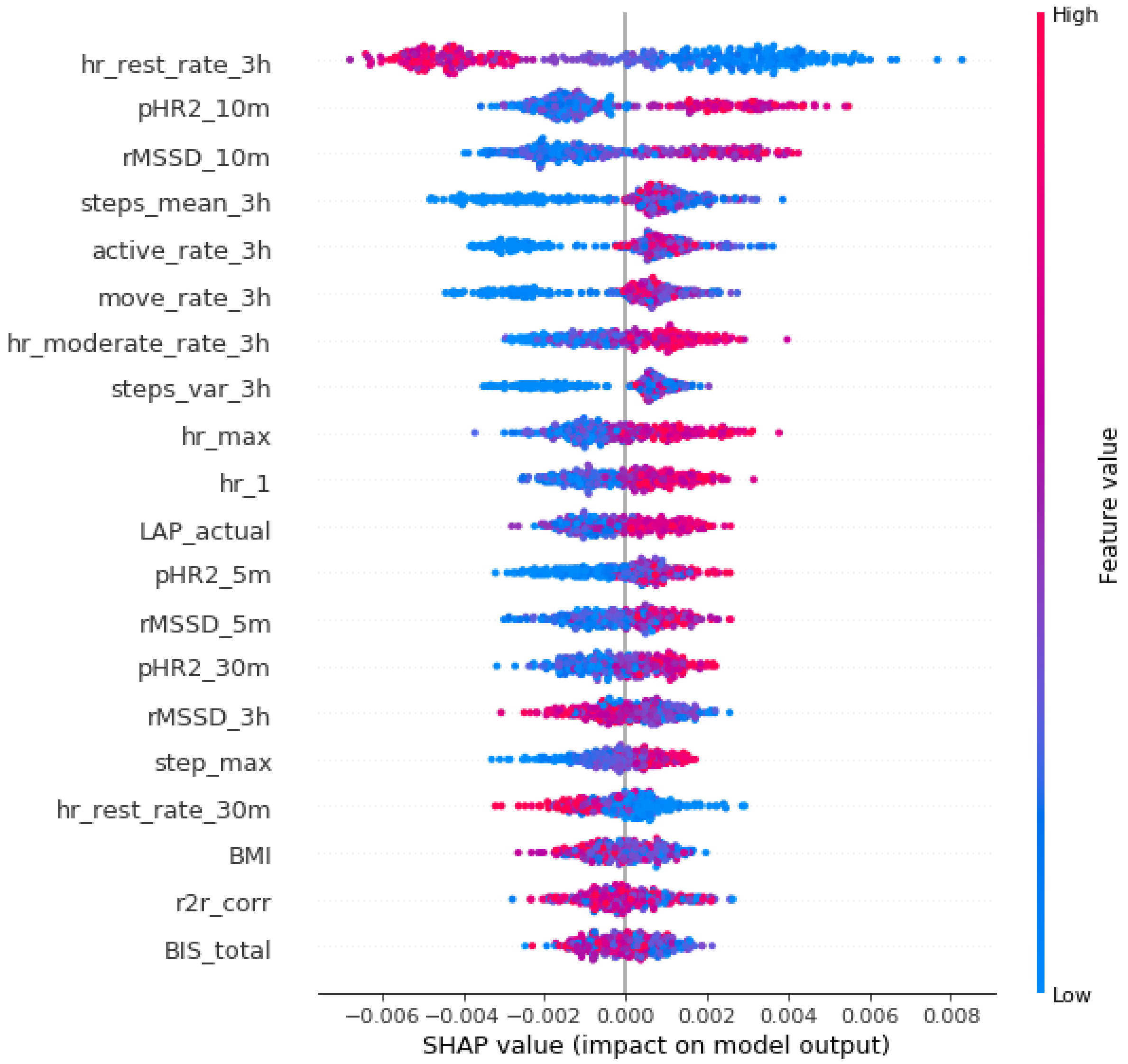


Fig 6. The Shapley Values of the features from the best model that predict relative valence.

There are three findings from Table 4 and Figure 6.

1) Using additional data (psychological, demographic data, physical health data) does help improve the prediction. In both scenarios, models trained with all data give better performance. F1 score improved by 0.109 and 0.086 for current users and new users respectively.
2) The precision of prediction for new users is lower than the precision of random guesses.
3) Fitbit data are very important. The top 10 most important features are all engineered from Fitbit data which represents the regulatory activity of heart rate or activity level.

Conclusions

1. In general, the data (wearable data, psychological data, health metrics and demographic data) would help the prediction of emotional states for both current users and new users.

As mentioned above, for positive/negative emotion prediction, the use of additional data helps improve the F1 scores by 0.119 and 0.104 for current users and new users respectively. For relative valence prediction, the use of additional data helps improve the F1 scores by 0.109 and 0.086 for current users and new users respectively.

2. Previous emotional states, psychological features, age and heart rate feature have a significant correlation on absolute level of valence. While wearable data reflects relative valence, especially heart rate related features.

Psychological features are one time measurements, they are highly correlated with the base level of emotional states for each subject. Therefore, when predicting the absolute level of valence, the positive/negative emotion, their influence is significant. In contrast, the relative valence is generated by comparing valence score with its mean within each subject. In this case, the influence of one time measurements are being cancelled out by the subtraction. Instead, Fitbit data stands out as it capture the regulatory activity and the change of activity, which are correlated to the fluctuation of emotional states.

3. It is hard to predict emotional states for a new user. The pattern we have learnt from current users could actually harm the prediction for new users.

In both cases, prediction for positive/negative emotion and prediction for relative valence, the precision of prediction for new users are worse than random guesses. This means the knowledge we extracted from current users could not be applied to new users. An possible explanation for this observation is that emotional characteristics differ by person, and it is hard to find common patterns with respect to emotional states. This explanation could also be backed up by figure 5, in which we have shown the widespread histogram of the mean and standard deviation of positive score and negative score, indicating the variability of emotional traits by person.

Limitations

1. Most of the subjects are between 20-30 and 50-60 years old. So our study applied to subjects within this age range.

2. The results are based on a small dataset (1872 observations), so the models suffer from errors caused by variation.

3. All the experiments are conducted in the US, and half of the participants are college students. So, the generalization capability of our models might be compromised.

References

1. Fredrickson, B. L. (2000). Cultivating positive emotions to optimize health and well-being. Prevention & treatment, 3(1), 1a.
2. EMOTION DETECTION AND RECOGNITION (EDR) MARKET - GROWTH, TRENDS, AND FORECAST (2019 - 2024)
3. Emotion Detection and Recognition Market by Software Tool (Facial Expression & Emotion Recognition, Gesture & Posture Recognition, and Voice Recognition), Application (Law Enforcement, Surveillance, & Monitoring; Entertainment & Consumer Electronics; Marketing & Advertising; and Others), Technology (Pattern Recognition Network, Machine Learning, Natural Language Processing, and Others), and End User (Commercial, Industrial, Defense, and Others) - Global Opportunity Analysis and Industry Forecast, 2017-2023
4. Levenson, R. W. (1992). Autonomic nervous system differences among emotions.
5. Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological psychology*, *84*(3), 394-421.
6. Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of emotions*, *2*, 173-191
7. Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Trans. Pattern Anal. Mach. Intell., 23, 1175–1191. https://doi.org/10.1109/34.954607
8. Choi, J., & Gutierrez-Osuna, R. (2009). Using Heart Rate Monitors to Detect Mental Stress. 2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 219–223.https://doi.org/10.1109/BSN.2009.13
9. de Santos Sierra, A., Sanchez Avila, C., Guerra Casanova, J., & Bailador del Pozo, G. (2011). A Stress-Detection System Based on Physiological Signals and Fuzzy Logic. IEEE Transactions on Industrial Electronics, 58(10), 4857–4865. https://doi.org/10.1109/TIE.2010.2103538
10. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., Alex, & Pentland. (2014). Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. Proceedings of the ACM International Conference on Multimedia - MM ’14, 477–486. https://doi.org/10.1145/2647868.2654933

11. Lawanont, W., Mongkolnam, P., Nukoolkit, C., & Inoue, M. (2018, June). Daily stress recognition system using activity tracker and smartphone based on physical activity and heart rate data. In International Conference on Intelligent Decision Technologies (pp. 11-21). Springer, Cham.
12. Liu, Z., Xu, A., Guo, Y., Mahmud, J. U., Liu, H., & Akkiraju, R. (2018, April). Seemo: A computational approach to see emotions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).
13. Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivee, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. Journal of Abnormal Child Psychology, 19, 285-300
14. Kobau, R., Sniezek, J., Zack, M. M., Lucas, R. E., & Burns, A. (2010). Well-being assessment: An evaluation of well-being scales for public health and population estimates of well-being among US adults. Applied Psychology: Health and Well-being, 2(3), 272-297. doi:http://dx.doi.org/10.1111/j.1758-0854.2010.01035.x
15. Cloninger CR (1987) The Tridimensional Personality Questionnaire, Version iv. St. Louis, MO: Department of Psychiatry, Washington University School of Medicine.
16. Lessing, E. E. (1968). Demographic, developmental and personality correlates of length of future time perspective (FTP). J. Personal. 36: 183-201.
17. Costa Jr, P.T., & McCrae, R.R. (1985).The NEO personality inventory manual. Odessa, FL: Psychological Assessment Resources.
18. Tsai, J.L. Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. Journal of Personality and Social Psychology, 90, 288-307.
19. Reise, S. P., Moore, T. M., Sabb, F. W., Brown, A. K., & London, E. D. (2013). The Barratt Impulsiveness Scale-11: reassessment of its structure in a community sample. Psychological assessment, 25(2), 631–642. doi:10.1037/a003216
20. Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. Journal of Personality and Social Psychology, 67, 319-333

Supplemental Material

Data cleaning and preprocessing

We joined all the tables using the `subject id` as a primary key in a wide-format data frame.. The researchers used instruments with different scales for each experiment. In order to solve this, we standardized the scales used for the variables NEO, AVI, BIS and FTP. Other steps included:

- Removed all the variables that have more than 140 missing values or more than 10% of missing ratio.
- Removed all the observations with more than 4 missing values.
- Removed all the observations without a `start survey` parameter in the experience sample table.

One hot encoded the following nominal variables

- Ethnicity
- Marital Status
- Sex
- Period_of_day

Encoded the following ordinal variables into an integer format

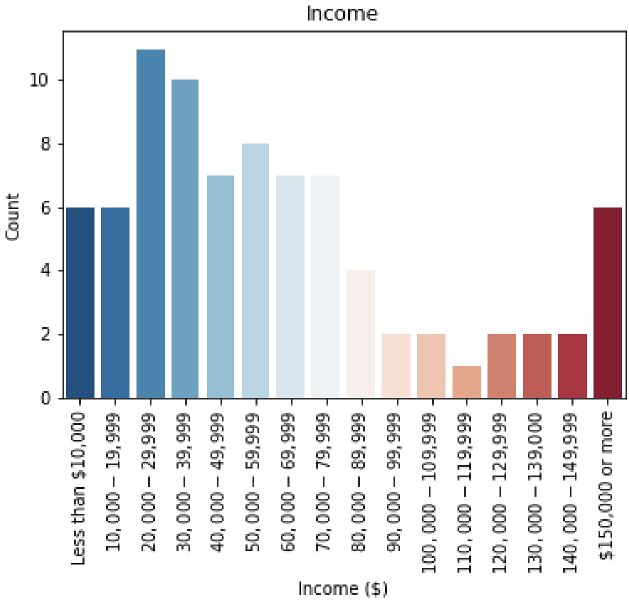
- Household_income

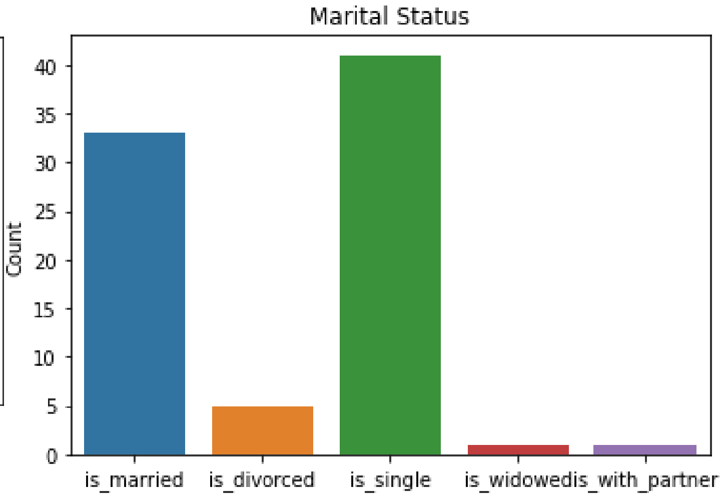
Type of data	**Variable name**	**Description**	**Measure Type**	**Sampling rate**	**Mean (SD)**
Psychometric	SBQ	Social Behavior Questionary [13]	Score	Constant	54.44 (4.59)
	SWLS	Satisfaction With Life Scale [14]	Score	Constant	20.32 (5.38)
	TPQ-NS	Trait Dimension Personality Questionary - Novelty Seeking [15]	Score	Constant	12.63 (4.61)
	FTP	Future time perspective [16]	Score	Constant	52.83 (9.35)
	NEO	Personality - Conscientiousness, Extraversion, Neuroticism [17]	Score	Constant	Supplement mat.
	AVI	Affect Valuation Index [18]	Score	Constant	Supplement mat.
	BIS	Impulsiveness [19]	Score	Constant	57.07 (9.83)
	BISBAS	Activation/Inhibition [20]	Score	Constant	Supplement

					mat.
Demographic	Age	Age measured in Years	Years	Constant	39.96 (16.45)
	Children	# of subject's children	# chidren	Constant	0.98 (1.36)
	Sex	Self-reported sex	Category	Constant	Female 42, Male 41
	Household Income	Self-reported total household income ($)	Ordinal	Constant	Fig below
	Marital Status	Self-reported marital status	Category	Constant	Fig belowt.
	BMI	Body Mass Index	BMI score	Constant	25.88 (4.83)
Heart Rate	Value	# heart palpitations per minute	Float	1 / 10 sec. (avg)	-
	Date and Time	Date and time where the value was recorded	Timestamp	1 / minute	-
Steps	Value	# of steps per minute	Float	1 / minute	-
	Date and Time	date and time where the value was recordeD	Timestamp	1 / minute	-

Table 5. Variables used in every complete data entry. Every observation in our models was composed by this data and the target variable(Positive Negative Emotion or Normalized Valence) in every case.

Histogram for the variables Income and Marital Status:





EDA

Correlations between labels

Correlations between labels make sense.
Positive measures and negative measures are positively correlated among each other respectively. And positive measures are negatively correlated with negative measures.

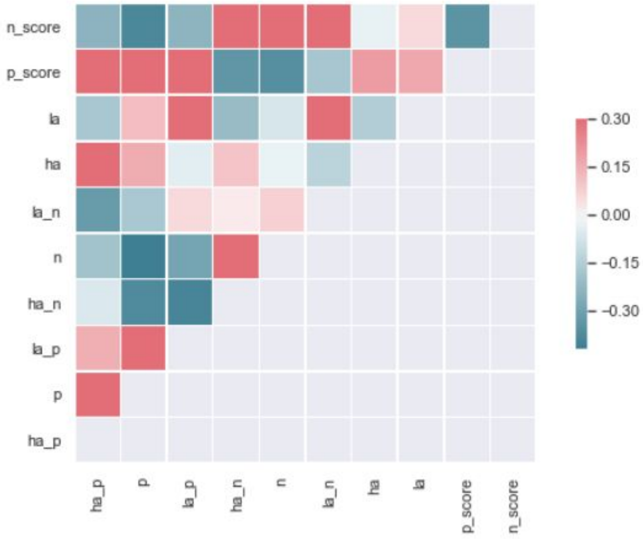


Fig 7. Correlations between emotional states.

Valence by Age

The distribution of subjects' age. Two age groups (20s & 50s) constitute ~70% of all subjects.

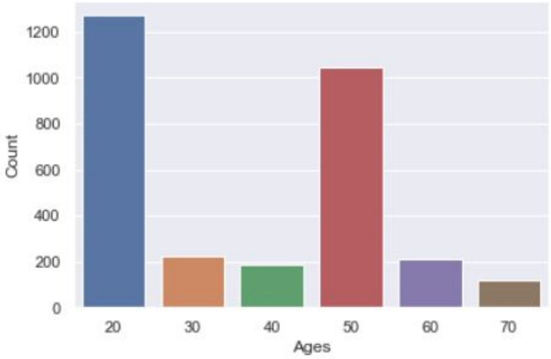


Fig 8. Age distribution of the participants.

Below are the distribution plots of p_score, n_score by age group. As we can see, the 50s group seem to be happier, and less unhappier than the 20s group. It might be because their emotional regulation skills are better.

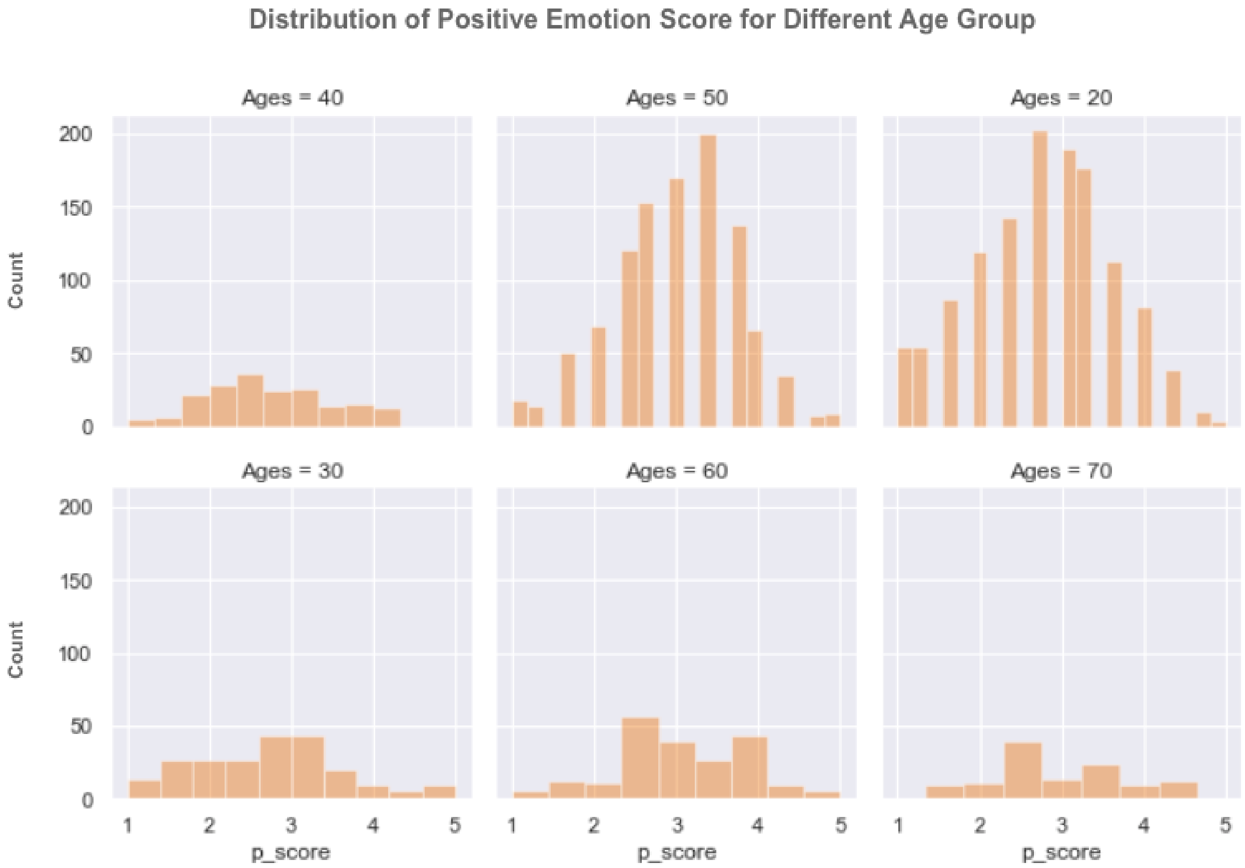


Fig 9. Distribution of Positive Emotion Score for different age groups.

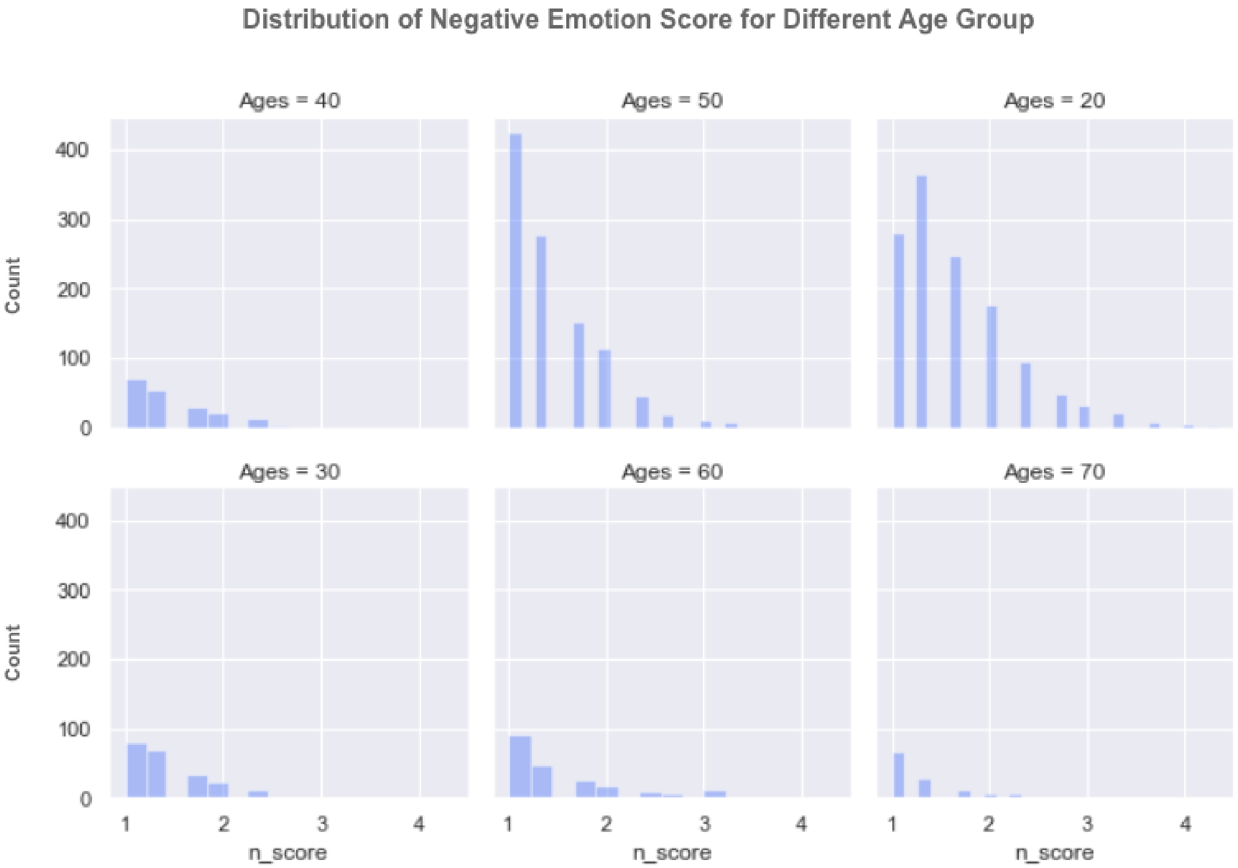


Fig 10. Distribution of Negative Emotion Score for different age groups.

Definition for Engineered Features

Feature Name	**Definition**
steps_max_<time window>	The maximum by minute step count within the time window
steps_min_<time window>	The minimum by minute step count within the time window
steps_mean_<time window>	The average by minute step count within the time window
steps_var_<time window>	The variance of by minute step count within the time window
steps_median_<time window>	The median of by minute step count within the time

	window
move_rate_<time window>	The percentage of the number of minutes within the time window with step count greater than 0
active_rate_<time window>	The percentage of the number of minutes within the time window with step count greater than 10
very_active_rate_<time window>	The percentage of the number of minutes within the time window with step count greater than 20
running_rate_<time window>	The percentage of the number of minutes within the time window with step count greater than 30
SDNN_<time window>	Standard deviation of heartbeat intervals within the time window. The heartbeat intervals are generated by the inverse of beats per minute
pHR2_<time window>	The percentage of the difference between adjacent HR greater than 2 within the time window
rMSSD_<time window>	Root of mean squared HR change within the time window
low_hr_<time window>	Lowest heart rate within the time window
high_hr_<time window>	Highest heart rate within the time window
l_h_<time window>	low_hr_<time window> / high_hr_<time window>
CR_<time window>	Highest HR within the time window / Highest HR so far
hr_mean_<time window>	The average heart beat per minute within the time window
hr_var_<time window>	The variance of the heart beat per minute within the time window
hr_std_<time window>	The standard deviation of the heart beat per minute within the time window
hr_median_<time window>	The median of the heart beat per minute within the time window
hr_rest_rate_<time window>	The percentage of the number of minutes within the time window with heart rate < 30 percentile heart rate of the previous time

hr_moderate_rate_<time window>	The percentage of the number of minutes within the time window with heart rate > 50 percentile heart rate of the previous time
hr_very_active_rate_<time window>	The percentage of the number of minutes within the time window with heart rate > 80 percentile heart rate of the previous time

Table 6. Engineered features used in our study. All these features are generated through the Fitbit activity data (step, heart rate). The step data contains step counts per minute, and the heart rate data contains heart rate count per minute. <time window> here refers to the time period within which we calculated the feature. In this study we used 5 time windows: 5 mins, 10 mins, 30 mins, 1 hr, 3 hrs.