RESEARCH

Predicting subjective well-being in a high-risk sample of Russian mental health app users

Polina Panicheva[1], Larisa Mararitsa[1,2], Semen Sorokin[1], Olessia Koltsova[1*] and Paolo Rosso[3,1]

[*]Correspondence: ekoltsova@hse.ru
[1]Laboratory for Social and Cognitive Informatics, HSE University, Russia
Full list of author information is available at the end of the article

Abstract

Despite recent achievements in predicting personality traits and some other human psychological features with digital traces, prediction of subjective well-being (SWB) appears to be a relatively new task with few solutions. COVID-19 pandemic has added both a stronger need for rapid SWB screening and new opportunities for it, with online mental health applications gaining popularity and accumulating large and diverse user data. Nevertheless, the few existing works so far have aimed at predicting SWB only in terms of Diener's Satisfaction with Life Scale. None of them analyzes the scale developed by the World Health Organization, known as WHO-5 – a widely accepted tool for screening mental well-being and, specifically, for depression detection. Moreover, existing research is limited to English-speaking populations, and tend to use text, network and app usage types of data separately. In the current work, we cover these gaps by predicting both mentioned SWB scales on a sample of Russian mental health app users who represent a population with high risk of mental health problems. In doing so, we employ a unique combination of phone application usage data with private messaging and networking digital traces from VKontakte, the most popular social media platform in Russia. As a result, we predict Diener's SWB scale with the state-of-the-art quality, introduce the first predictive models for WHO-5, with similar quality, and reach high accuracy in the prediction of clinically meaningful classes of the latter scale. Moreover, our feature analysis sheds light on the interrelated nature of the two studied scales: they are both characterized by negative sentiment expressed in text messages and by phone application usage in the morning hours, confirming some previous findings on subjective well-being manifestations. At the same time, SWB measured by Diener's scale is reflected mostly in lexical features referring to social and affective interactions, while mental well-being is characterized by objective features that reflect physiological functioning, circadian rhythms and somatic conditions, thus saliently demonstrating the underlying theoretical differences between the two scales.

Keywords: digital traces; subjective well-being; mental health prediction

Introduction

In recent years, evaluation, analysis and improvement of subjective well-being (SWB) has gained a growing attention of both researchers and practitioners (1; 2). Attention to SWB has naturally been coupled with the increasing research interest in depression - the leading cause of disability and subjective well-being loss worldwide (3; 4). The COVID-19 pandemic, resulting in the shift to hybrid work and the decline in face-to-face communication has put many individuals at additional mental health risks (5; 6). Some of the most widely available instruments to mitigate such risks are online and mobile services that offer quick screening tests of subjective well-being and mental health states and automatically generate respective recommendations. More than 240 mental health apps are available in the App Store today, some of which are extensively using machine learning for classifying and scoring their users in terms of their psychological or mental conditions (7–9). These apps attract users concerned with their psychological conditions who voluntarily donate parts of their digital traces; thus, these apps become natural hubs accumulating data on individuals at risk. Such data, if available, may open wide possibilities for the development of open source algorithms for early automatic detection of threats to well-being in high-risk populations with their digital traces.

Subjective well-being (SWB) is most commonly defined in accordance with Diener's approach (10) as a person's satisfaction with their life (which constitutes SWB's cognitive component) and the prevalence of positive emotions over negative ones (affective balance, which constitutes SWB's affective component). To date, about 100 assessment tools measuring about 200 facets of well-being have been proposed, thus complicating the selection of relevant metrics (1). The two most widely used SWB measurement tools are Diener's Satisfaction with Life Scale (SWLS) (10) and the scale introduced by the World Health Organization in 1998, known as the WHO-5 index (11). The former aims to capture generalized long-term subjective well-being, while the original goal of the latter was to screen, diagnose and rate depression. Later, Bech, one of the WHO-5 developers, also showed that this scale is equally good at detecting high degrees of psychological well-being, which he proposed to consider a component of mental health, along with the absence of depression symptoms (12).

Both SWLS and WHO-5 are short unidimensional 5-item scales with proven validity and reliability (α coefficients 0.79-0.89 for the former and 0.82-0.95 for the latter) (13–15). Both have become common for well-being screening in a wide range of populations and among different nationalities (15–18). The wide use and the proven quality of these metrics defines their choice for our research in automatic SWB prediction; however, some more details on their distinctive features should be added.

SWLS, apart from being centered on pleasure and satisfaction, is also meant to be time- and dimension-independent. The first feature means that it is not tied to a specific time interval and measures satisfaction with our past, present and future. The second feature refers to the generalized character of such satisfaction, not being tied to any particular dimension of human life, such as health, relationships or finance. The choice of the dimensions to be taken into account and the weight assigned to them is left with the subject and is expected to be based on a blend of objective reality and the subject's subjective experience of it. It is assumed that a person is able to adequately assess her well-being and has all the necessary and unbiased information for that (10).

SWLS is widely used by psychologists, public health professionals, and economists. According to the World Happiness Report, SWLS provides a more informative measure for international comparisons of well-being than some measures capturing affective component only (19). Importantly, SWLS is stable under unchanging conditions, but is sensitive to changes in life circumstances, such as marriage or childbirth that increase SWLS score, and job loss or relocation that decrease it (20). It is also predictive of physical and physiological outcomes, as judged from a 4-year follow-up period in the same study.

In contrast to SWLS, WHO-5 index aims at a brief assessment of emotional well-being over a 14-day period (thus containing no cognitive component and being time-sensitive). Its items represent positive affect whose absence corresponds to the depression symptoms (negative affect). This is an important advantage of WHO-5 as the subjects are not forced to confess of the presence of any unpleasant and potentially hard-to-admit negative emotions or states. WHO-5 has been proven effective both for depression detection (21; 22) and for measuring the effects of treatments on well-being across multiple patient groups (18). Being a short, sensitive, specific and non-invasive tool, it gains over more detailed, but heavier methods for preliminary depression and suicide risk assessment in settings without psychological/psychiatric expertise. WHO-5 has been adopted in various research fields such as suicidology, geriatrics, youth problems and alcohol abuse studies, personality disorder research, and occupational psychology (15; 23).

Thus, WHO-5 and SWLS, being psychometrically sound screening tools with known outcomes, also measure complementary aspects of subjective well-being. Although measures of emotional affect and reported life satisfaction often correlate, substantial divergences have been found. For instance, almost half of the people who rated themselves as ‘completely satisfied’ also reported significant symptoms of anxiety and distress (17). Therefore, quality of life in the current coronavirus crisis is usually measured with both scales (5; 6; 24–26): while WHO-5 helps to assess influence of different practices on SWB and the persistence of diminished well-being beyond and during COVID-19, SWLS shows how people feel and how their life perspective changes due to the pandemic. This complementarity indicates the importance of comparative research in prediction of both metrics.

Such taks is novel for SWB prediction with digital traces: despite the advances in detection of specific mental health problems and the attempts to predict some SWB metrics, no research so far has been dedicated to predicting WHO-5 and its comparison with SWLS in terms of digital behavior traces; moreover, most of the research is limited to English-speaking populations. Best models predicting SWLS with digital traces from social media, search engine and smartphone activity data demonstrate performance below 0.4 in terms of Pearson correlation – a well-known threshold for correlation between psychological characteristics and objective behavior (27; 28) (see also (29; 30) for an overview). None of the models combines language, social media and smartphone usage data.

The goal of this study is to predict individual WHO-5 and SWLS levels with a new combination of digital traces in a high risk Russian-speaking population, to find out which features are the most predictive and what the overall predictive power of our models is. We thus address a completely novel task of comparative prediction of two different aspects of subjective well-being, which should have different objective indicators and suggest different actions to be taken by the user. Additionally, as we find that certain levels of WHO-5 and SWLS indices are themselves predictive of depression, we predict such thresholds as well.To do so, we make use of a sample of 372 psychological application users that allows us to predict SWB of high-risk individuals in real-world conditions with their private messages, social media data and mobile device usage traces. For this goal, we use extensive feature engineering combined with regression and classification modeling, the first type of models aimed at SWB score prediction, and the second – and depression risk identification based on theoretically justified thresholds. We also check our regression models against newest neural network approaches that, however, do not show sufficient quality at the dataset of our size.

The rest of the paper is structured as follows. In the next section we review the existing literature in prediction of SWB and related psychological and mental health phenomena with digital traces. Next, we describe our dataset, our numerous features and the approach to their engineering, as well as the models used. In the Results section we report our best models' performance and the most useful features. In the Discussion section we interpret our results and indicate the most important limitations. We conclude with the perspectives for future research. ...

Subjective well-being prediction

Prediction of internal psychological and mental states from objective behavior features is a highly difficult task (28; 31). Psychological theory views such states as latent constructs that are not expected to fully correlate with any observable patterns since the former are not thought of as reducible to the latter in principle. This may be one of the reasons why such correlation is seldom high, although this is a subject for further research. As both high SWB and the absence of mental disorder symptoms have been shown to be components of mental health (12; 32) prediction of both SWB and mental disorder constitutes two related tasks. However, due to the different nature of these two concepts, the former is usually evaluated with continuous predictive models, while the detection of the latter is most often formulated as a classification task.

Detection of mental disorders

A vast amount of studies predict specific mental health conditions with digital traces, mostly with the data from social media, such as Facebook and Twitter. The most widely analyzed conditions of such studies are depression and Post Traumatic Stress Disorder (PTSD) (33–37). Other conditions include Bipolar Disorder, Anxiety and Social Anxiety Disorder, eating disorders, self-harm and suicide attempt (38–41). Linguistic features used typically include word n-grams, sentiment, specific lexica (e.g., LIWC) and topic modelling, with other features related to social networks, emotions, cognitive styles, user activity and demographics (33–38; 41). Model evaluation metrics include Area Under the Curve (AUC), Precision, Accuracy of classification, and Correlation for continuous measurements. The results for binary mental health problem identification are high, reaching an AUC of 0.7-0.89, Precision up to 0.85, and Accuracy of 0.69-0.72 (29).

Ground truth information in such studies is obtained from different sources, leading to different quality. Most studies use either self-reported survey data (33; 36) or self-declared mental illness (35; 38). The latter is prone to errors and bias induced by specific data collection methods.

In a recent study Eichstaedt et al. (37) effectively predict depression of Facebook users against medical records information. The authors use a 6-month history of

Facebook statuses posted by 683 hospital patients, of whom 114 were diagnosed with depression (rate similar to the general population), and classify depression VS other medical diagnoses with an AUC = 0.72. Features of Facebook statuses include words and word bigrams, temporal characteristics of posting activity, metainformation on post length and frequency, topics and dictionary categories, with interpersonal, emotional and cognitive categories being among the best predictors.

The effects of smartphone usage on mental disorders, until very recently, have been mostly studied with self-reported data (see (42; 43) for an overview). Meanwhile, smartphone apps that collect usage data provide an unprecedented opportunity to access objective and precise information on smartphone application usage. Hung et al. (44) find that phone call duration and rhythm patterns are predictive of negative emotions, while Saeb et al. (45) predict depressive symptom severity with geographical location and phone usage frequency information. However, as feature engineering with phone app usage data requires considerable time and effort (46), the potential of such data of psychological research is yet to be discovered.

Prediction of SWB levels

There have been a few studies aimed at predicting subjective well-being levels, mostly with regression, which obtain modest results. Individual and relational well-being was predicted from social network data (27; 47) and from objective smartphone use data (48). The reported results are close to the upper bound expected in this task: the meta-analytic correlation between digital traces and psychological well-being has been estimated as r = 0.37 across nine studies, including prediction of subjective well-being, emotional distress and depression (27). The only study that reaches a higher correlation of 0.66 in one of the models (48) does not specify the scales used for measuring SWB; however, interestingly, it finds that while some apps predictably have a negative effect on well-being, others affect it positively.

Diener's SWLS, to our knowledge, has been predicted in only four studies that use digital traces in a cross-validated setting. In his pioneering study, Kosinski et al. (49) predicted SWLS with linear regression for 2,340 Facebook users based on 58K 'Likes' - preferences of webpages indicated by the users. The Likes data dimensionality was reduced to top 100 components in a SVD model based on a larger dataset (58K users). The obtained correlation reached r = 0.17, whereas empirical test-retest correlation for SWLS was r = 0.44.

Collins et al. (50) predicted SWLS with Random Forest Regression and various Facebook features, including demographics, networking data, photos, likes, ground truth Big Five traits of the users, of their significant others and friends, and predicted Big Five as a proxy. The best result for a sample of 1,360 users with Big Five features as a proxy reached the Mean Absolute Error (MAE) = 0.162, whereas the model with social network features produced MAE = 0.173 for SWLS. Unfortunately, no other evaluation metrics were reported in this study. Schwartz et al. (51) applied Ridge Regression to predict SWLS of 2,198 individuals using their Facebook statuses. Thousands of linguistic features were extracted from the status texts, including 2,000 topics obtained with LDA, word uni- and bi-grams, LIWC and sentiment lexica. A message-user level cascaded aggregation model was additionally trained on a disjoint dataset, which allowed to improve regression results from Pearson r = 0.301 to r = 0.333. Facebook status data were also used by Chen et al. (52) to predict SWLS of 2,612 users. Features included affect measured by sentiment word usage, 2K LDA-based topics and 66 LIWC categories. After feature selection with Elastic Net regression, Random Forest model was tested for prediction of an unseen subset. The results reach RMSE 1.30 (0.217 when rescaled to [0;1]) and r = 0.36.

There is a certain number of studies predicting SWB with app usage data. Some of them use self-reported measures of app use (53), while others collect objective data (48; 54). Correlation in David's model range from 0.31 to 0.66, however, the research does not specify the scales used for measuring SWB. At the same time, interestingly, it finds that while some apps predictably have a negative effect on well-being, others affect it positively. Gao and colleagues (54) report correlation from 0.34 for male users to 0.66 for female users in the task of predicting SWLS, however, they do not report the full feature set and the contribution of each feature in their best models. Instead, they mention that the most predictive variables are communication apps, certain types of games and the frequency of photo taking. None of these studies mentions cross-validation.

Overall, although the results of subjective well-being prediction are promising, several gaps in the existing research can be identified. First, WHO-5, which is an effective screening tool for various mental health conditions and subjective well-being, has never been studied in a predictive research design. Second, all the studies predicting SWLS are limited to English-speaking populations and respective linguistic features. Moreover, these works only address Facebook digital traces, including profile, texts and likes. Finally, only scarce feature interpretation is reported in the previous studies, and digital trace manifestations of different well-being dimensions have never been compared.

Our approach

In this study, we set out to predict two different concepts of subjective well-being: one combining affective balance and life satisfaction (measured by SWLS index and further referred to as satisfaction-related SWB) and the other conceptualized as a reflection of mental health (measured by WHO-5 index and further referred to as mental SWB). We perform our prediction on the texts of private messages, social media and smartphone usage information. We show that in a sample of participants who have completed both the WHO-5 and one of the questionnaires on mental health well-being, including depression, anxiety and stress, effective cutoff thresholds for WHO-5 values can be chosen to address all of these conditions in terms of high sensitivity and specificity. [1] [2] We combine social media and phone app usage data to generate features predictive of SWB, and perform regression and classification experiments in a cross-validated Machine Learning design. The novelty of the current study lies in the following:

1 We present **the first study so far on predicting subjective well-being measured by WHO-5**, including classification, allowing us to predict the risk of a variety of mental health conditions with highly promising results;

2 We use a dataset of a psychological application users, allowing us to predict **subjective well-being in real-world conditions for a sample with high mental risks**, which has never been done before;

3 To our knowledge, our study is the first to address subjective well-being prediction in a **Russian-speaking population** and respective data: the Russian social network VKontakte and texts in the Russian language;

4 This is the first study to **combine** language, social media and phone app usage features in well-being research.

5 We are the first to **compare satisfaction-based and mental SWB**, analyzing their intersections and differences in terms of predictive features.

Materials and methods

Dataset

Our dataset was collected in collaboration with Humanteq social analytics company, using its DigitalFreud app (DF) – a phone application for psychological self-assessment promoted among Android-based smartphone users through Google Ads. Users were offered to take as many free tests as they wanted (including personality traits, cognitive, motivation and SWB tests) and to explicitly consent to the access to their VKontakte profile data and / or smartphone use data. Based on the test results, users were offered psychological feedback and analytics on the use of VK and / or their smartphones. Privacy policy included a clause stating that the data could be used for research. The study was approved by the HSE Ethics Committee. Nevertheless, the data were anonymized prior to the analysis. No personal information (i.e. allowing to identify the users) was included in the sample. In particular, all the user profile ids were encrypted.

The initial sample included 2,050 accounts of DF users who have completed at least one of the two questionnaires of our interest: SWLS (10) or WHO-5 (55).

The following digital traces data were available for the participants:

- DF profile data;
- VKontakte user data;
- Phone application data.

As most of our data in the sample is sparse, our **final sample** used in prediction contains digital traces by 372 users. The procedure of data cleaning that produced this dataset is given in Appendix 1 .The dataset is small also due to the fact that the data on both well-being questionnaires combined with personal digital traces is highly difficult to obtain, as it requires both considerable effort from a user on completing the questionnaires, and trust allowing them to share sensitive digital traces. However, our dataset is uniquely tailored to the task of predicting SWB in a high-risk population of mental health app users.

Additionally, there is a **heldout dataset**, which consists of messages written by 572 users, who lack other important features for prediction (demographics, phone app usage) but have text data. The **heldout dataset** is used for preliminary feature selection (see sections *Words, Word clusters* below). Before feature selection, texts were tokenized with *happiestfuntokenizing*[1] and lemmatized it with *pymorphy* (56).

The **phone app dataset** consists of phone application usage data by 992 users who lack other important features for prediction. The **phone app dataset** was used for preliminary phone application categorization and feature engineering.

[1] https://github.com/dlatk/happierfuntokenizing.

We also collected a sub-sample of users (N = 417), who have completed the WHO-5 and at least one of the following questionnaires measuring mental health conditions (**mental health dataset**):

1 Depression measured with PHQ-9 (57);
2 Anxiety measured with GAD (58);
3 Stress: the Perceived Stress Scale (PSS) (59) [ ].

The **mental health dataset** was used in the WHO-5 classification task to select cutoff thresholds of the classes to be predicted, so the former would be representative of a range of mental health conditions. Overall, both SWB levels measured by SWLS and WHO-5 were predicted with regression and classification.

Self-reported well-being measures

Satisfaction-related well-being scale (SWLS). The SWLS questionnaire was translated and adapted to Russian by Ledovaya et al. (60).

The questionnaire contains 5 statements, each characterized by 7-point Likert scale ranging from 1 (strongly agree) to 7 (strongly disagree). The resulting SWLS score ranges from 5 (low satisfaction) to 35 (high satisfaction). In our sample, 1,727 accounts have information about the SWLS score.

Mental well-being scale (WHO-5). We use the Russian-language version of WHO-5 scale developed by WHO itself (55). Each of WHO-5 items is scored on a 6-point Likert scale ranging from 0 (at no time) to 5 (all of the time). The WHO-5 score ranges from 5 (absence of well-being) to 30 (maximal well-being). In our sample, 1,791 accounts have information about the WHO-5 score.

Mental well-being classes. As mentioned earlier, WHO-5, unlike SWLS, is indicative of a range of mental health conditions (23) and was directly designed to detect one of them (11). Decisions of mental health, be it screening test results or medical diagnoses, are usually binary indicating the absence or the presence of a disease. For such tasks scales need to be transformed into sets of discrete classes based on a certain threshold values. To choose optimal cutoff values for our classification task, we analyzed the **mental health dataset** of 417 DF users who have completed both WHO-5 and one of the mental conditions questionnaires. We tried our different WHO-5 thresholds to reach better sensitivity and specificity in representing the following conditions: PHQ/GAD = 10 for depression and anxiety (61), and PSS = 21 for stress [et al. 2016]. Additionally, as from our earlier work (62) we know that classes derived from scale reduction might be better predicted in a trinary design in social science NLP tasks, we also experimented with three-class divisions.

Eventually, our analysis resulted in the following cutoff values of the normalized WHO-5 scale:

- Binary cutoff = 0.51 with classes containing 221 and 151 users in the low and high SWB classes, respectively;
- Trinary cutoffs = [0.35; 0.59] with classes containing 111, 158 and 103 users in the low, medium and high SWB classes.

Table 1 illustrates sample statistics for each of the mental health conditions, and specificity and sensitivity in terms of the selected WHO-5 cutoff values.

Table 1 Specificity and sensitivity of the selected WHO-5 cutoff values in the mental health dataset.

Condition	N (mental health dataset)	Metric	Binary cutoff (0.51)	Lower trinary cutoff (0.35)	Upper trinary cutoff (0.59)
Depression	344	Sensitivity	0.80	0.49	0.90
		Specificity	0.58	0.87	0.45
Anxiety	309	Sensitivity	0.82	0.53	0.92
		Specificity	0.54	0.83	0.41
Stress	323	Sensitivity	0.85	0.47	0.93
		Specificity	0.66	0.88	0.50

In our high-risk sample of mental health app users, the binary WHO-5 cutoff value 0.51 allows to reach high sensitivity across the analyzed mental health conditions, while preserving moderate specificity. The trinary cutoff values 0.35 and 0.59 allow to obtain low and high mental well-being classes with very high specificity and sensitivity, respectively, across the mental health conditions.

Digital traces

DigitalFreud profile Account information about the DF user includes encrypted DF and Vkontakte user ids, SWLS and WHO-5 scores, gender, birth year, education, employment and marital status, and date and time of the DF app installation.

Vkontakte user information We use the following data from Vkontakte social network API:

1 User Profile data. Although Vkontakte (VK) API provides access to potentially rich user information, in practice users seldom fill in their profiles, and the data is sparse. As a result, we only use gender, birthdate, and the number of friends and subscriptions in our analysis.
2 Wall posts (text, date and time, information on reposting with the original post contents and encrypted user id, number of reposts, comments and likes) available for 1,871 users.
3 Directed private messages (text, date and time, encrypted author and addressee ids) available for 1,044 users.

Humanteq chooses to match DF data with VK data because VK is the most popular social networking site in Russia and, additionally, it provides access to the widest range of social media data.

Phone application usage Phone application usage was monitored for one week following the initial consent obtained from the user when she started using DF, which was consistent both with the app's terms of use and the policies of the Android platform. The collected information includes name and package of the application, start time and duration of the application usage in foreground in milliseconds. It is available for 992 users. In a few cases when the users quit the phone app data sharing before the end of the week, the recorded period was shorter.

Descriptive statistics

The main parameters of the descriptive statistics for our **final dataset** of 327 users are given in Tables 2 and 3. Consistent with Collins et al (50), we normalize both well-being scores to the ranges between [0,1]; to do so, we subtract 5 from both scores, then multiply SWLS values by 1/30, and WHO-5 values by 1/25. Additionally, the distribution of the SWB and demographic data in the **final dataset** is illustrated in Appendix 2, Figures 1-4.

Table 2 Descriptive statistics for subjective well-being, age and gender in the final dataset.

	N	Range	Mean	Std	Mean (norm)	Std (norm)	Cronbach's α
SLWS	372	5 - 35	18.30	6.73	0.4433	0.2243	0.8365
WHO-5		5 - 30	16.51	4.66	0.4604	0.1865	0.8205
Age		18 - 53	23.06	5.06			
Gender		Male, Female	298 (80%) Female				

SWLS and WHO-5 intercorrelate strongly with r = 0.568, p < 10-32. The level of internal consistency of both scales is high (Cronbach's $\alpha > 0.82$).

Table 3 Descriptive statistics for the textual and phone app usage features in the final dataset.

Data	Sum	Mean	Median	Min	Max
Messages	6,739K	18,115	10948.5	52	131,368
Message alters	53K	143	107.5	2	1,029
Message volume (chars)	160,707K	432,009	240,831	671	2,983,231
Posts	7K	19	4	0	1880
Post volume (chars)	857K	2,303	84	0	87,708
App Usage (seconds)	1,573K	4,231	3,715.5	24	16,329

Compared to some other data on subjective well-being in Russia (63), where WHO-5 mean was 0.60 +- 0.191, SWB in our final sample is lower (0.46 +- 0.187), while male and older participants exhibit higher scores in both studies. The lower SWB levels in our sample are explained by self-selection of specific individuals to DF app: it naturally attracts users interested in seeking psychological and mental health information and advice, i.e., potentially more likely to have problematic mental health conditions. This is consistent with the age-gender distribution of our sample: as female and younger individuals prove to have lower SWB in other studies, our dataset is predictably skewed towards containing more females (80%) and young people (mean age 23 +- 5 y.o). The bias of our dataset towards lower levels of SWB is, however, in line with our research goal of studying high-risk populations.

Feature engineering

For our task of SWLS and WHO-5 prediction, we construct features of three main types:

- User metadata and overall activity: demographics, DF & VK profile statistics, and overall phone app usage statistics;
- Textual, or linguistic features:
  - Words;
  - Sentiment scores;
  - RuLIWC;
  - Word clusters;
- Phone app usage statistics by app category.

Overall, we constructed 660 for SWLS and 651 for WHO-5. Most features were calculated as counts, ratios or counts by time period directly from the **final dataset**.

However, words and word clusters as features were trained on the **heldout dataset** that does not intersect with the **final dataset**. Of these features, only those that correlated with the target variables were selected for the main experiments. In the main experiments, the features were submitted to the regression or classification models, which performed on the **final dataset** that was divided into train, development and test subsets in 10-fold cross-validation scenario. In this scenario, (1) multiple models were trained on the train set, (2) recursive feature elimination was performed on the development set based on MAE of the models, and (3) final scores for each feature type and each model were computed based on the test set. More details on the main experiment procedure are given in the *Machine Learning Experiments* section.

User metadata and overall activity features

There are 40 features describing demographics, overall activity patterns based on DF and VK profile data and overall phone application usage data (see Table 4). In building phone app usage features, we follow the previous research (64; 65) which identified three- and six-hour periods of online activity to be significant markers of mental illness. In our research, we break phone app usage into three-hour periods of activity. Some features have been excluded from the analysis, as they contained sparse data.

Table 4 User metadata and overall activity features.

Feature name	Description	Number
Age	-	1
Gender	-	1
NVkFriends	№ of friends in VK	1
AllAlters	№ of alters (accounts that a user has a message history with) in the last 12 months	1
Subscriptions	№ of VK pages subscriptions	1
Mess_ 1	Total number of messages written in the last 30 days	1
MessChars_ 1	Total size (in characters) of messages written in the last 30 days	1
growth-2to-1weighted	Weighted difference between total size of messages written in the months -1 and -2	1
altersdiff	Weighted difference between numbers of alters in the months -1 and -2	1
AppUsage1Week	Number of active app usage instances in the period of app data sharing time	1
AllAppTime1Week	Total time of phone app usage in the period of app data sharing time (seconds)	1
RatioAppTime1Week	Ratio of phone app usage time in the week of app data sharing time	1
AppUsage 0-3, 3-6, 6-9, 9-12, 12-15, 15-18, 18-21, 21-24	Total time of phone app usage in 3-hour time periods	8
AppUsage 0-3, 3-6, 6-9, 9-12, 12-15, 15-18, 18-21, 21-24 Ratio	Ratio of phone app usage time in 3-hour time periods normalized by total app usage time	8
Alters-1 - -12	Number of alters in every month (30 days) before the DF install time, for months between -1 and -12	12
	Total	40

Linguistic features

Our extensive analysis of user texts has shown that VK public wall posts are too sparse and include mostly web link content, which does not allow for effective prediction. As a result, we construct all the linguistic features based on private messages written by the users in VK messenger, mostly - during the year preceding the installation of DF app.

Sentiment scores. We use six features representing the proportions of positive and of negative words in the messages created during the last month or the last year prior to the data collection, our in the entire messaging history of a user.

Words. We adopt the open-vocabulary approach to word features predictive of well-being (66). Given the small size of our final dataset (372 observations), using all the frequent words as features (12K words with frequency >=200) would inevitably result in overfitting. To overcome this and to select a reasonable number of interpretable features, we use the heldout dataset as follows:

- First, a sub-sample of users who have filled both well-being questionnaires was selected from the heldout dataset (396 users);
- Next, we selected 12.5K words occurring more than 200 times in the joint one-year long message collection of all users and calculated their TfIDF scores using 396 individual message collections as 396 texts for such calculation;
- We filtered out words with p > 0.01 in the ANOVA tests relating these words to SWLS and WHO-5 values in the heldout dataset, which has resulted in the selection of 165 words for SWLS and 224 words for WHO-5 (see Appendix 3 for the full list). Words belonging to either of these sets (353 words) are used as features for prediction.

RuLIWC. For obtaining closed-vocabulary features, we used RuLIWC dictionary - a translation of the most prominent categories of the Linguistic Inquiry and Word Count (LIWC, (67)) performed by Panicheva & Litvinova (68). RuLIWC consists of eight word categories: Bio, Cognitive, Social, Time, Percept and subcategories of the latter: Feel, Hear, See, with 563-2,624 words in each category and 20-303 words in each sucategory. For this research, RuLIWC feature values have been computed as the sums of all the words' TfIDF values for every user. All the words regardless of their (in)frequency were accounted for.

Word clusters. Content features were computed by clustering words with a word2vec semantic model (69) based on the **heldout dataset**. The word2vec model we used had been trained on the web-based Taiga corpus containing over 5 billion words (70) by Kutuzov & Kuzmenko (71), with skipgram algorithm, vector dimensionality = 300, and window size = 2. For clustering, we used 7,128 words present in the model vocabulary with frequency >= 200 in the **heldout dataset**. Next, we performed KMeans clustering with cosine distance and 300 clusters. As KMeans algorithm is stochastic and may give very different results in different runs, we used the following procedure to obtain reproducible cluster solutions:

- We employed cluster regularization, where the regularization parameter was the sum of p-values of the cluster occurrence correlation with SWLS or WHO-5[2]; the regularization weights were [0; 50; 100; 500];
- For every weight value, ten random cluster solutions were obtained;

[2] https://arxiv.org/abs/1804.10742, the code https://github.com/Kipok/clr_prediction was modified and applied.

- Based on these solutions, consensus cluster solutions were constructed[3] with the following thresholds: [0.25, 0.45, 0.65, 0.75, 0.85];
- This resulted in five consensus cluster solutions for every weight value, thus the overall number of solutions totaling to 20.
- In each solution, clusters were additionally augmented with infrequent words in the dataset, every infrequent word being ascribed to the closest cluster. Thus each of 20 solutions was supplemented by a paired solution with augmented clusters.

The clustering results were evaluated on the **heldout dataset** as follows:

- For every cluster solution, only the clusters correlated with $p < 0.05$ with SWLS or WHO-5 on the were used as features;
- Each cluster feature was computed as the sum of the respective words' TfIDF values;
- The resulting features were used for RandomForest regression predicting SWLS and WHO-5 on the **heldout dataset**, with 10-fold train/test cross-validation and recursive feature elimination;
- The best cluster features were chosen by Mean Average Error (MAE) of the regression models trained on the **heldout dataset**; later they were used for prediction on the **final dataset**.

The main parameters of the resulting feature sets are described in Table 5.

Table 5 Best word cluster features.

	Regularization weight	Consensus clustering threshold	Infrequent words	No of clusters	MAE
SWLS	500	0.45	-	28	0.1704
WHO-5	0	0.45	+	19	0.1525

Phone app categories and usage features

App categories, or types were obtained from the **phone app dataset** data by using 53 app categories generated automatically from 28K app descriptions and by manually uniting them into larger groups as described in (46; 48). As a result, we identified the following nine app categories: *Game, Education+Productivity, Tools, Entertainment, Personalization, Health+Medical, Social+Communication+Dating, Photography*, covering 21.5K apps, with the rest 6.5K apps having been assigned to *Other*. The main app usage features were calculated as the total time devoted to a certain app category (e.g. *Game, Photography* or *Other*) in each of eight three-hour time slots of a day, averaged over all days of a given user (9*8 = 72 features), as well as overall time spent for this category in the entire app usage history of an individual (9 features). Next, we constructed several normalized versions of each feature. Namely, we normalized them by the total app usage time in this category, and by the total app usage logged in the current three-hour period. This resulted in 9 + 72*3=225 features. The phone app category features are exemplified in Table 6.

[3] https://naeglelab.github.io/OpenEnsembles/_modules/finishing.html#majority_vote

Table 6 Phone app category features.

Feature type	№ of features	Example feature name	Description
Total time logged in category by a user	9	GAME	Total time logged in Game apps by a user
Total time logged in category in time period by a user	72	GAME_21-24	Total time logged in Game apps between 21 and 24h by a user
Total time logged in category in time period/total time logged in category by a user	72	PHOTOGRAPHY_0-3 / PHOTOGRAPHY	Ratio of time logged in Photography apps between 0 and 3 AM to total time logged in Photography apps by a user
Total time logged in category in time period/total time logged in time period by a user	72	EDUCATION + PRODUCTIVITY_15-18 / 15-18	Ratio of time logged in Education+Productivity apps between 15 and 18h AM to total time logged in apps between 15 and 18h AM by a user

Machine learning experiments

We performed specific experiments for each of our two subtasks: prediction of satisfaction-related and mental well-being scales and prediction of classes in both types of SWB. As we aimed at interpretable results, our main experiments were based on classical regressions. Simultaneously, to make sure that we obtain the best possible prediction quality with the available contemporary methods, we also carried out extensive experiments employing deep learning approaches (described in Appendix 4). However, they yielded inferior results. The two main possible reasons for that are the following (1) our data are hard to obtain, and the obtained data are sparse and loosely intersect between users, which reduces the sample significantly; (2) our message data is hierarchically organized, with numerous message alters for every participant, numerous messages addressing every alter, while additionally the number of alters and messages highly varies between the participants/alters (see Table 3 above).

Our experiment on prediction of SWLS and WHO-5 scales was performed using a 10-fold cross-validation design with train, development and test sets (298/37/37 users, 80/10/10%). The non-overlapping train, development and test sets were constructed as follows:

1 The sample was shuffled and sorted by the well-being values;

2 The sorted sample was divided into 10 bins containing 37 users each so that bin_i consisted of users with $index = i + K * 37$, where K varied in the range $[0; 36]$. Thus every bin was equally distributed in terms of the SWB values.

3 For i-th cross-validation fold, bin_i was used as the test set, bin_{i+1} - as the dev set, and the remaining users belonged to the training set.

Our evaluation metrics for **regression** include Mean Absolute Error (MAE), Pearson r and $R2$-score. Hyperparameter values were chosen inside the cross-validation loop based on the results obtained from development by MAE values. Recursive Feature Elimination (RFE) was performed based on the development set to identify the informative features in each cross-validation fold. RFE was adopted based on the earlier experiments which had shown the increase in model performance with RFE. Additionally, RFE allows to select a small number of informative features, improving the model interpretability. The selected best hyperparameters and features were used to evaluate the quality of prediction on the test set inside the cross-validation loop. In the end, the evaluation metrics were averaged across all 10 folds.

Predictions of SWLS and WHO-5 scores were performed with seven regression models, including Linear Regression with various regularization techniques, Decision Tree, and two ensemble methods (see Appendix 5). WHO-5 classification was performed with three classification models based on our preliminary experiments (Appendix 6).

Classification of individual WHO-5 levels was performed in a **binary** mode with two classes (*low VS high well-being*) and in a trinary mode with three classes (*low VS medium VS extremely high*). The models and hyperparameter values are described in Appendix 6. We report F1-macro and F1-weighted metrics over all the classes, as well as F1 metric for the lowest and the highest classes separately. We additionally report True Positive and False Positive Rates for the low well-being class, as these measures are typically used for screening test of various mental health conditions (cf. (37)).

All the calculations were performed in *python* with *pandas*, *scipy*, and *scikit-learn* libraries.

Results

Prediction of well-being scale values

The continuous modeling results for the SWLS and WHO-5 well-being values are presented in Tables 7 and 8, respectively.

Table 7 SWLS value prediction results.

Features	Best model		Results	
		MAE	Pearson R	R-2
Mean baseline		0.1853	-	-
Median baseline		0.185	-	-
Words	ElasticNet	0.1744	0.3402	0.1022
RuLIWC	DecisionTree	0.182	0.2168	0.0142
AppCats	ElasticNet	0.1762	0.2737	0.0172
Behavior	DecisionTree	0.1785	0.191	0.0195
Clusters	RandomForest	0.1814	0.1709	0.026
Clusters + AppCats + Behavior + Words	**ElasticNet**	**0.1698**	**0.4024**	**0.1045**
Clusters + AppCats + RuLIWC + Behavior + Words	**ElasticNet**	**0.1681**	**0.3776**	**0.1164**

The results for every individual feature set, and for the best feature sets in terms of every evaluation metric are included; the best results are highlighted in bold. The full results for all the feature set combinations are presented in Appendices 7, 8.

Table 8 WHO-5 value prediction results.

Features	Best model		Results	
		MAE	Pearson R	R-2
Mean baseline		0.1542	-	-
Median baseline		0.1533	-	-
Words	Lasso	0.1441	0.3179	0.0817
RuLIWC	Lasso	0.1529	0.1276	0.0197
AppCats	ElasticNet	0.1511	0.2172	0.0329
Behavior	DecisionTree	0.1497	0.2463	0.0096
Clusters	Lasso	0.1516	0.1533	0.0241
Clusters + RuLIWC + Words	AdaBoost	**0.1436**	**0.3202**	**0.081**
AppCats + RuLIWC + Behavior + Words	**ElasticNet**	**0.1438**	**0.367**	**0.1193**

Overall, the best feature set is words written by the users in messages, and the best model is ElasticNet.

Prediction of WHO-5 classes

The main classification results for the WHO-5 well-being are presented in Table 9. The full WHO-5 classification results are presented in Appendix 9.

Table 9 Best WHO-5 classification results.

Classification	Threshold	N (Classes)	Best model	Best features	F1-macro	F1-weighted	F1-low	F1-high	True Positive Rate (low)	False Positive Rate (low)
binary	0.51	221 / 151	AdaBoost	Words + RuLIWC + AppCats	0.692	0.706	0.768	0.616	0.792	0.404
binary majority baseline					0.378	0.456	0.373	0	1.0	1.0
trinary	0.35 / 0.59	111 / 158 / 103	AdaBoost	Clusters + RuLIWC + Words	0.483	0.493	0.502	0.433	0.450	0.161
trinary majority baseline					0.199	0.253	-	-	0.0	0.0

Significant features

The features in the best performing continuous models of satisfaction-related well-being (SWLS) and mental well-being (WHO-5) scales are illustrated in Tables 8 and 9. Only the features which were selected by RFE in at least five out of ten cross-validation folders are included; the features significant in both SWLS and WHO-5 regression are highlighted in bold. All the significant features are listed in Appendices 10, 11.

Discussion

In this paper, we have introduced a novel task of predicting mental well-being measured by WHO-5 index, as compared to traditionally studied satisfaction-related SWLS, with digital traces, and performed it in both continuous modeling and classification designs. In the latter, we have shown that the selected WHO-5 thresholds are representative of a range of three mental well-being-related conditions (depression, anxiety and stress) with high sensitivity and specificity. Furthermore, the results obtained in mental well-being classification are highly promising (0.792 True Positive Rate and 0.404 False Positive Rate) in the binary task with our highly sensitive threshold. This result is similar to the performance of the best existing models that predict other mental conditions with digital traces (29; 37). Likewise, our results of SWLS and WHO-5 scale prediction, with Pearson r = 0.402 and 0.367, respectively, improve the state-of-the-art metrics reported previously in similar tasks with cross-validation designs (50; 52). Since, as mentioned earlier, prediction of internal states with observable behaviors has its limitations (28; 29), the obtained correlation may be considered high. As a result, we obtain a model which is highly sensitive and sufficiently specific for identifying low levels of subjective well-being requiring intervention in a high-risk population of mental health application users. Our model is unique in the field of mental health prediction from digital traces, as it allows an overall screening for mental health risks, not limited to specific conditions reported in previous studies (see (27; 29; 47) for an overview).

We have performed a unique comparison of regression models predicting both SWLS and WHO-5 indices on the same sample. Our best models for both indices

Table 10 Predictive features in SWLS scale. Slang, misspellings and unconventional word forms are shown with an asterisk (*). Errors in lemmatization are enclosed in brackets.

Feature type	Feature	Translation / Description	Coefficient
Words	спать_[NOUN]	sleep_VERB	41086
	интим_NOUN	intimacy_NOUN (suggestive of 'intercourse')	-44937
	орг_NOUN*	org(aniser)_NOUN	23978
	дропнуть_VERB*	quit_VERB	-64677
	тратиться_VERB	spend_VERB	-24593
	отл_UNKN*	fine_UNKN	34184
	пояснение_NOUN	explanation_NOUN	-22499
	стебать_VERB*	bully_VERB (rude)	-28898
	[вифя]_NOUN*	wifi_NOUN	-48114
	спойлерить_VERB*	spoil_VERB	-48530
	ооохнуть_VERB*	gasp_VERB	-44864
	милый_COMP	nice_COMPARATIVE	56128
	[пиздёжа]_NOUN*	lie_NOUN (rude)	-22727
	обжечь_VERB	burn_VERB	-40019
Sentiment	**Negative_month**	negative sentiment in the last month	-29
Activity	**AppUsage9-12Ratio**	Ratio of phone app usage time between 9 and 12 AM normalized by total app usage time	10
	AppUsage0-3Ratio	Ratio of phone app usage time between 0 and 3 AM normalized by total app usage time	-8
AppCats	SOCIAL + COMMUNICATION + DATING_0-3 / SOCIAL + COMMUNICATION + DATING	Ratio of time logged in Social + Communication + Dating apps between 0 and 3 AM to total time logged in Social + Communication + Dating apps	11
	PHOTOGRAPHY_18-21/18-21	Ratio of time logged in Photography apps between 18 and 21h PM to total time logged in apps between 18 and 21h PM	8

show similar performance in terms of correlation and R2 metrics, but WHO-5 is predicted better in terms of MAE across all feature combinations; however, this is likely an outcome of different distributions of SWLS and WHO-5 in our sample (see Fig. 1,2, Table 1 above).

Our design also allows us to compare the features predictive of life satisfaction-related SWB and mental SWB. Although our experiments have revealed only two highly predictive features that are common for both SWLS and WHO-5, they are highly interpretable in terms of psychological theory. These two metrics are (1) phone app usage time between 9 and 12 AM normalized by total app usage time, and (2) negative sentiment expressed in private messages in the last month, which have positive and negative coefficients, respectively, in both SWLS and WHO-5 tasks. Both of these findings confirm previous results obtained in various populations: participants affected by depression and other low SWB conditions have been found less likely than average individuals to participate in online activities in the morning hours around 9-10 AM (64; 65), while their circadian rhythms are often disrupted (7). Such disruption is what usually accompanies insomnia or hypersomnia, a symptom of the major depressive disorder listed in DSM-5 (72), the

Table 11 Predictive features in WHO-5 scale.

Feature type	Feature	Translation / Description	Coefficient
AppCats	GAME_3-6/GAME	Ratio of time logged in Game apps between 3 and 6h AM to total time logged in Game apps	-5
	ENTERTAINMENT_3-6/ENTERTAINMENT	Ratio of time logged in Entertainment apps between 3 and 6h AM to total time logged in Entertainment apps	4
	HEALTH+MEDICAL_3-6/HEALTH+MEDICAL	Ratio of time logged in Health+Medical apps between 3 and 6h AM to total time logged in Health+Medical apps	3
	PERSONALIZATION_0-3 / 0-3	Ratio of time logged in Personalization apps between 0 and 3h AM to total time logged in apps between 0 and 3h AM	-4
	EDUCATION + PRODUCTIVITY_9-12 / EDUCATION + PRODUCTIVITY	Ratio of time logged in Education+Productivity apps between 9 and 12h AM to total time logged in Education+Productivity apps	-3
	TOOLS_18-21 / 18-21	Ratio of time logged in Tools apps between 18 and 21h PM to total time logged in apps between 18 and 21h PM	-3
	SOCIAL + COMMUNICATION + DATING_3-6 / SOCIAL + COMMUNICATION + DATING	Ratio of time logged in Social+Communication+Dating apps between 3 and 6 AM to total time logged in Social+Communication+Dating app	7
	GAME_9-12/GAME	Ratio of time logged in Game apps between 9 and 12h AM to total time logged in Game apps	2
	OTHER_3-6/OTHER	Ratio of time logged in Other apps between 3 and 6h AM to total time logged in Other apps	-2
	ENTERTAINMENT_9-12 / ENTERTAINMENT	Ratio of time logged in Entertainment apps between 9 and 12h AM to total time logged in Entertainment apps	2
	PHOTOGRAPHY_0-3 / PHOTOGRAPHY	Ratio of time logged in Photography apps between 0 and 3h AM to total time logged in Photography apps	-2
	EDUCATION + PRODUCTIVITY_21-24 / EDUCATION + PRODUCTIVITY	Ratio of time logged in Education+Productivity apps between 21 and 24h PM to total time logged in ducation+Productivity apps	-2
RuLIWC	Bio_RuLIWC	Words related to Biological processes in RuLIWC	-20
Words	😘_UNKN	😘_emoji	35
	но_CONJ	but_CONJ	-16
Activity	**AppUsage9-12Ratio**	Ratio of phone app usage time between 9 and 12 AM normalized by total app usage time	7
Sentiment	**Negative_month**	negative sentiment in the last month	-33
	Negative_year	negative sentiment in the last year	-29
	Negative_all	negative sentiment in overall messages	-23

Diagnostic and Statistical Manual of Mental Disorders developed by the American Psychological Association.

Negative sentiment has been shown to correlate negatively with life satisfaction (33; 52; 73) and subjective well-being (63). Negative sentiment in written or oral speech may also sometimes, although not always, be a manifestation of depressed mood, another symptom of depressive disorder according to DMS-5.

Thus, these two highly predictive features intersecting in both SWLS and WHO-5 prediction models can indicate different degrees of SWB: from simple dissatisfaction with life, circumstances or personal achievements (relevant for SWLS), to a deterioration in mental or physical condition and serious symptoms of the depressive spectrum (relevant for WHO-5). They can be recommended for use across various SWB-prediction tasks.

Predictors unique for satisfaction-related well-being are much more dominated by verbal features related to affect-laden psychological and social content. They are often obscene lexemes, but also represent both negative and positive sentiment polarities (*quit_VERB, spend_VERB, fine_UNKN, explanation_NOUN, bully_VERB, spoil_VERB, gasp_VERB, nice_COMPARATIVE*). Association of positive lexica with SWB is consistent with Weismayer (74), who also finds negative relation of SWB with lexica expressing anger and fear. Some of our predictive words are likely to express these emotions (e.g. *bully [rude], burn, lie [rude], gasp*). Also, these lexica fit well with some of the ontologies developed for depression detection (44). Prevalence of lexical features among SWLS predictors suggests that this index, indeed, captures subjective perception of well-being rather than symptoms of mental disorders, such as depression.

On the contrary, in mental well-being level prediction phone app usage features take a clear lead, especially those related to the ratio of nighttime app usage (3-6 AM). Additionally, lexica related to biological processes are also a distinctive marker of low WHO-5 levels. All this aligns well with the primary goal of this index to reveal depression and its proved ability to differentiate between problematic mental health states and high levels of mental health-related well-being. Specifically, app usage rhythms and biological lexica are likely to be manifestations of such depression symptoms as increase or decrease in either weight or appetite, insomnia or hypersomnia, and fatigue or loss of energy (75). At the same time, they can be markers of a poor physical condition, which is also detected by WHO-5 (18). Finally, significance of negative sentiment in long periods of messaging (1 year and longer) for WHO-5 levels suggests that mental SWB measured by this index might in fact have a more stable behavioral pattern than SWLS. This contradicts the goals of both WHO-5 and SWLS and thus requires further examination.

Conclusions

The growing interest in tracking human mental states and in the development of mindfulness leads to the growth of applications that screen or even diagnose mental conditions and offer solutions for their correction, including those based on objective data. Our research has shown that it is possible to create machine learning models based on interpretable traits and predict various aspects of subjective well-being at the state-of-the-art level.

In doing so, we have performed **the first study on predicting subjective well-being measured by WHO-5**. We have demonstrated that certain WHO-5 level thresholds are indicative of a range of mental health conditions prevalent in a sample characterized by high risk of mental health problems. We have obtained promising results in classification of mental SWB into classes constructed based on these thresholds. This approach has allowed us to identify individuals affected by low subjective well-being with high recall and reasonable false positive rates, based on their digital traces.

Our study is also **the first to compare prediction performance and predictive features of mental SWB and satisfaction-related SWB**. We show that several predictors are shared by well-being measured by both WHO-5 and SWLS, and these digital traces are bluntly indicative of overall (un)well-being. At the same time, digital traces distinguishing between WHO-5 and SWLS are closely related to the conceptual difference between these two indices: while SWLS is characterized by expressions denoting affect-laden psychological and social content, WHO-5 levels are manifested in objective features reflecting physiological functioning and somatic conditions, i.e., lexica related to biological processes and circadian rhythm-related ratios of phone app usage.

To our knowledge, this is **the first approach to subjective well-being prediction in a Russian-speaking population**, and **the first to combine language, social network and phone app usage features** in well-being research. By leveraging phone app usage logs, profile and message data from the Russian social network VKontakte, we have been able to improve prediction of satisfaction-related SWB (SWLS) and propose a first predictive model for mental SWB (WHO-5). At the same time, as our sample has been very small and limited to a high-risk population, the study needs replication on larger samples representative of wider social and psychological groups. The major obstacle to this is that VK private message data are no longer available for any type of download, while other social media are even more restrictive. Development of public policies and regulations encouraging private data-collecting companies to share portions of their data for public good purposes is highly recommended.

Ethical statement

The study has been approved by the Higher School of Economics Committee on Interuniversity Surveys and Ethical Assessment of Empirical Research.

Declarations

Availability of data and materials

The data that support the findings of this study belong to the Humanteq company and were collected under specific terms of use. When onboarding in the Digital Freud app, users agreed to a privacy policy that explicitly prohibited data transfer to third parties, largely because the amount of data for each user does not allow completely anonymizing the dataset and contains sensitive information. Therefore restrictions apply to the availability of these data, which is why they are not publicly available. A fraction of the data can however be obtained from the authors upon reasonable request and with permission of the Humanteq company.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PP: participated in hypothesis formulation, engineered all features, ran the main predictive models and prepared the initial draft.

LM: collected all the data, formulated the initial research problem and edited the manuscript from the psychological point of view.

SS: ran all experiments with neural networks and prepared the Appendix.

OK: participated in hypothesis formulation, curated dataset formation and feature engineering, ran several exploratory models, provided major editing of the final draft.

PR: provided the overall research design, suggested methods and approaches, participated in editing the final draft.

Acknowledgements

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE University) in 2021.

This work is the result of the collaboration with Paolo Rosso in the framework of his virtual online internship at the National Research University Higher School of Economics due to COVID-19.

This research was supported in part through computational resources of HPC facilities at NRU HSE.

Author details

[1]Laboratory for Social and Cognitive Informatics, HSE University, Russia. [2]Humanteq, Russia. [3]Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, Spain.

References

1. Linton, M.-J., Dieppe, P., Medina-Lara, A.: Review of 99 self-report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. BMJ open **6**(7), 010641 (2016)
2. Goodday, S.M., Geddes, J.R., Friend, S.H.: Disrupting the power balance between doctors and patients in the digital era. The Lancet Digital Health **3**(3), 142–143 (2021)
3. Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J.: Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. The lancet **367**(9524), 1747–1757 (2006)
4. Barzilay, R., Moore, T.M., Greenberg, D.M., DiDomenico, G.E., Brown, L.A., White, L.K., Gur, R.C., Gur, R.E.: Resilience, covid-19-related stress, anxiety and depression during the pandemic in a large population enriched for healthcare providers. Translational psychiatry **10**(1), 1–8 (2020)
5. Wilke, J., Hollander, K., Mohr, L., Edouard, P., Fossati, C., González-Gross, M., Sánchez Ramírez, C., Laiño, F., Tan, B., Pillay, J.D., *et al.*: Drastic reductions in mental well-being observed globally during the covid-19 pandemic: results from the asap survey. Frontiers in medicine **8**, 246 (2021)
6. Pieh, C., Budimir, S., Delgadillo, J., Barkham, M., Fontaine, J.R., Probst, T.: Mental health during covid-19 lockdown in the united kingdom. Psychosomatic medicine **83**(4), 328–337 (2021)
7. Rohani, D.A., Faurholt-Jepsen, M., Kessing, L.V., Bardram, J.E.: Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. JMIR mHealth and uHealth **6**(8), 165 (2018)
8. Devakumar, A., Modh, J., Saket, B., Baumer, E.P., De Choudhury, M.: A review on strategies for data collection, reflection, and communication in eating disorder apps. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–19 (2021)
9. Huang, Y.-N., Zhao, S., Rivera, M.L., Hong, J.I., Kraut, R.E.: Predicting well-being using short ecological momentary audio recordings. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–7 (2021)
10. Diener, E., Emmons, R.A., Larsen, R.J., Griffin, S.: The satisfaction with life scale. Journal of personality assessment **49**(1), 71–75 (1985)
11. Organization, W.H., et al.: Wellbeing measures in primary health care: the depcare project: report on a who meeting. Stockholm, Sweden, 12–13 (1998)
12. Bech, P., Olsen, L.R., Kjoller, M., Rasmussen, N.K.: Measuring well-being rather than the absence of distress symptoms: a comparison of the sf-36 mental health subscale and the who-five well-being scale. International journal of methods in psychiatric research **12**(2), 85–91 (2003)
13. McDowell, I.: Measures of self-perceived well-being. Journal of psychosomatic research **69**(1), 69–79 (2010)
14. Diener, E., Inglehart, R., Tay, L.: Theory and validity of life satisfaction scales. Social Indicators Research **112**(3), 497–527 (2013)
15. Sischka, P.E., Costa, A.P., Steffgen, G., Schmidt, A.F.: The who-5 well-being index–validation based on item response theory and the analysis of measurement invariance across 35 countries. Journal of Affective Disorders Reports **1**, 100020 (2020)
16. Downs, A., Boucher, L.A., Campbell, D.G., Polyakov, A.: Using the who-5 well-being index to identify college students at risk for mental health problems. Journal of College Student Development **58**(1), 113–117 (2017)
17. Kusier, A.O., Folker, A.P.: The well-being index who-5: hedonistic foundation and practical limitations. Medical humanities **46**(3), 333–339 (2020)
18. Kusier, A.O., Folker, A.P.: The satisfaction with life scale: Philosophical foundation and practical limitations. Health Care Analysis **29**(1), 21–38 (2021)
19. Helliwell, J.F., Layard, R., Sachs, J., De Neve, J.-E.: World Happiness Report 2020. Sustainable Development Solutions Network, New York (2020)
20. Luhmann, M., Lucas, R.E., Eid, M., Diener, E.: The prospective effect of life satisfaction on life events. Social Psychological and Personality Science **4**(1), 39–45 (2013)
21. Blom, E.H., Bech, P., Högberg, G., Larsson, J.O., Serlachius, E.: Screening for depressed mood in an adolescent psychiatric context by brief self-assessment scales–testing psychometric validity of who-5 and bdi-6 indices by latent trait analyses. Health and quality of life outcomes **10**(1), 1–6 (2012)
22. Krieger, T., Zimmermann, J., Huffziger, S., Ubl, B., Diener, C., Kuehner, C., Holtforth, M.G.: Measuring depression with a well-being index: further evidence for the validity of the who well-being index (who-5) as a measure of the severity of depression. Journal of affective disorders **156**, 240–244 (2014)
23. Topp, C.W., Østergaard, S.D., Søndergaard, S., Bech, P.: The who-5 well-being index: a systematic review of the literature. Psychotherapy and psychosomatics **84**(3), 167–176 (2015)
24. Chouchou, F., Augustini, M., Caderby, T., Caron, N., Turpin, N.A., Dalleau, G.: The importance of sleep and physical activity on well-being during covid-19 lockdown: reunion island as a case study. Sleep medicine **77**, 297–301 (2021)
25. Brindal, E., Ryan, J.C., Kakoschke, N., Golley, S., Zajac, I.T., Wiggins, B.: Individual differences and changes in lifestyle behaviours predict decreased subjective well-being during covid-19 restrictions in an australian sample. Journal of Public Health (Oxford, England) (2021)
26. Gierc, M., Riazi, N.A., Fagan, M.J., Di Sebastiano, K.M., Kandola, M., Priebe, C.S., Weatherson, K.A., Wunderlich, K.B., Faulkner, G.: Strange days: Adult physical activity and mental health in the first two months of the covid-19 pandemic. Frontiers in Public Health **9**, 325 (2021)
27. Settanni, M., Azucar, D., Marengo, D.: Predicting individual characteristics from digital traces on social media: A meta-analysis. Cyberpsychology, Behavior, and Social Networking **21**(4), 217–228 (2018)
28. Meyer, G.J., Finn, S.E., Eyde, L.D., Kay, G.G., Moreland, K.L., Dies, R.R., Eisman, E.J., Kubiszyn, T.W., Reed, G.M.: Psychological testing and psychological assessment: A review of evidence and issues. American psychologist **56**(2), 128 (2001)
29. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences **18**, 43–49 (2017)
30. Novikov, P., Mararitsa, L., Nozdrachev, V.: Inferred vs traditional personality assessment: are we predicting the same

thing? arXiv preprint arXiv:2103.09632 (2021)

31. Guntuku, S.C., Lin, W., Carpenter, J., Ng, W.K., Ungar, L.H., Preoţiuc-Pietro, D.: Studying personality through the content of posted and liked images on twitter. In: Proceedings of the 2017 ACM on Web Science Conference, pp. 223–227 (2017)
32. Bech, P.: Subjective positive well-being. World Psychiatry **11**(2), 105 (2012)
33. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
34. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: Clpsych 2015 shared task: Depression and ptsd on twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 31–39 (2015)
35. Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H.A., Ungar, L.: The role of personality, age, and gender in tweeting about mental illness. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 21–30 (2015)
36. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3187–3196 (2015)
37. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences **115**(44), 11203–11208 (2018)
38. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 51–60 (2014)
39. Coppersmith, G., Ngo, K., Leary, R., Wood, A.: Exploratory analysis of social media prior to a suicide attempt. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pp. 106–117 (2016)
40. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 152–162 (2017)
41. Uban, A.-S., Chulvi, B., Rosso, P.: An emotion and cognitive based analysis of mental health disorders from social media data. Future Generation Computer Systems (2021)
42. Lee, Y.-K., Chang, C.-T., Lin, Y., Cheng, Z.-H.: The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress. Computers in human behavior **31**, 373–383 (2014)
43. Sheldon, P., Rauschnabel, P., Honeycutt, J.M.: The Dark Side of Social Media: Psychological, Managerial, and Societal Perspectives. Academic Press, ??? (2019)
44. Hung, G.C.-L., Yang, P.-C., Chang, C.-C., Chiang, J.-H., Chen, Y.-Y.: Predicting negative emotions based on mobile phone usage patterns: an exploratory study. JMIR research protocols **5**(3), 160 (2016)
45. Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., Mohr, D.C.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. Journal of medical Internet research **17**(7), 175 (2015)
46. Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T., *et al.*: Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National Academy of Sciences **117**(30), 17680–17687 (2020)
47. Luhmann, M.: Using big data to study subjective well-being. Current Opinion in Behavioral Sciences **18**, 28–33 (2017)
48. David, M.E., Roberts, J.A., Christenson, B.: Too much of a good thing: Investigating the association between actual smartphone use and individual well-being. International Journal of Human–Computer Interaction **34**(3), 265–275 (2018)
49. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the national academy of sciences **110**(15), 5802–5805 (2013)
50. Collins, S., Sun, Y., Kosinski, M., Stillwell, D., Markuzon, N.: Are you satisfied with life?: Predicting satisfaction with life from facebook. In: International Conference on Social Computing, Behavioral-cultural Modeling, and Prediction, pp. 24–33 (2015). Springer
51. Schwartz, H.A., Sap, M., Kern, M.L., Eichstaedt, J.C., Kapelner, A., Agrawal, M., Blanco, E., Dziurzynski, L., Park, G., Stillwell, D., *et al.*: Predicting individual well-being through the language of social media. In: Biocomputing 2016: Proceedings of the Pacific Symposium, pp. 516–527 (2016). World Scientific
52. Chen, L., Gong, T., Kosinski, M., Stillwell, D., Davidson, R.L.: Building a profile of subjective well-being for social media users. PloS one **12**(11), 0187278 (2017)
53. Linnhoff, S., Smith, K.T.: An examination of mobile app usage and the user's life satisfaction. Journal of strategic marketing **25**(7), 581–617 (2017)
54. Gao, Y., Li, H., Zhu, T.: Predicting subjective well-being by smartphone usage behaviors. In: HEALTHINF, pp. 317–322 (2014)
55. Индекс общего (хорошего) самочувствия/ВОЗ (вариант 1999 г.). https://www.psykiatri-regionh.dk/who-5/Documents/WHO5$_Russian.pdf$
56. Korobov, M.: Morphological analyzer and generator for russian and ukrainian languages. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 320–332 (2015). Springer
57. Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9: validity of a brief depression severity measure. Journal of general internal medicine **16**(9), 606–613 (2001)
58. Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B.: A brief measure for assessing generalized anxiety disorder: the gad-7. Archives of internal medicine **166**(10), 1092–1097 (2006)
59. Cohen, S., Kamarck, T., Mermelstein, R., *et al.*: Perceived stress scale. Measuring stress: A guide for health and social scientists **10**(2), 1–2 (1994)
60. Ledovaya, Y.A., Bogolyubova, O.N., Tikhonov, R.V.: Stress, well-being and the dark triad. Psikhologicheskie Issledovaniya **8**(43), 5 (2015)
61. Spitzer, R., Williams, J., Kroenke, K.: Instruction manual: Instructions for patient health questionnaire (phq) and gad-7 measures. PHQ and GAD-7 instructions (1990)

62. Pronoza, E., Panicheva, P., Koltsova, O., Rosso, P.: Detecting ethnicity-targeted hate speech in russian social media texts. Information Processing & Management **58**(6), 102674 (2021)
63. Bogolyubova, O., Panicheva, P., Ledovaya, Y., Tikhonov, R., Yaminov, B.: The language of positive mental health: Findings from a sample of russian facebook users. SAGE Open **10**(2), 2158244020924370 (2020)
64. Birnbaum, M.L., Wen, H., Van Meter, A., Ernala, S.K., Rizvi, A.F., Arenare, E., Estrin, D., De Choudhury, M., Kane, J.M.: Identifying emerging mental illness utilizing search engine activity: A feasibility study. Plos one **15**(10), 0240820 (2020)
65. Ten Thij, M., Bathina, K., Rutter, L.A., Lorenzo-Luaces, L., van de Leemput, I.A., Scheffer, M., Bollen, J.: Depression alters the circadian pattern of online activity. Scientific reports **10**(1), 1–10 (2020)
66. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., *et al.*: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), 73791 (2013)
67. Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K.: The development and psychometric properties of liwc2015. Technical report (2015)
68. Panicheva, P., Litvinova, T.: Matching liwc with russian thesauri: An exploratory study. In: Conference on Artificial Intelligence and Natural Language, pp. 181–195 (2020). Springer
69. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
70. Shavrina, T., Shapovalova, O.: To the methodology of corpus construction for machine learning:«taiga» syntax tree corpus and parser. Proceedings of the "Corpora, 78–84 (2017)
71. Kutuzov, A., Kuzmenko, E.: Webvectors: a toolkit for building web interfaces for vector semantic models. In: International Conference on Analysis of Images, Social Networks and Texts, pp. 155–161 (2016). Springer
72. American Psychiatric Association, D., Association, A.P., et al.: Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American Psychiatric Association (2013)
73. Wang, N., Kosinski, M., Stillwell, D., Rust, J.: Can well-being be measured using facebook status updates? validation of facebook's gross national happiness index. Social Indicators Research **115**(1), 483–491 (2014)
74. Weismayer, C.: Investigating the affective part of subjective well-being (swb) by means of sentiment analysis. International Journal of Social Research Methodology, 1–16 (2020)
75. Fried, E.I., Nesse, R.M.: Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. Journal of Affective Disorders **172**, 96–102 (2015)