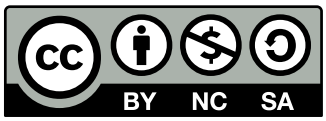
Aalto University
School of Electrical
Engineering

Master's programme in Computer, Communication and Information Sciences

Depression and Suicide Risk Detection From Internet Usage Traces

Emilia Marchese

Master's Thesis
2024



Aalto University
School of Electrical
Engineering

Author Emilia Marchese

Title Depression and Suicide Risk Detection From Internet Usage Traces

Degree programme Computer, Communication and Information Sciences

Major Machine Learning, Data Science and Artificial Intelligence

Supervisor Prof. Juhi Kulshrestha

Advisor Prof. Talayeh Aledavood

Date 9 February 2023 **Number of pages** 92+13 **Language** English

Abstract

Depression is one of the leading causes of illness and disability worldwide. Over the past few decades, there has been a surge in our reliance on the internet, advancing the prospect of utilizing our online behavior as a diagnostic tool for identifying depression and suicide risk. At the same time, it raises the question of potential associations between internet use and mental health. Previous research on internet usage for mental health assessment pertains mainly to data from mobile devices from small homogeneous populations. This thesis explores the potential of internet usage (IU) features from desktop and mobile devices for depression and suicide risk assessment using a large heterogeneous population of about 900 individuals per device type.

This study shows that IU features can distinguish people with no depression symptoms from people with high depression severity with an accuracy of 0.61, which improves to 0.66 when combined with sociodemographic features. The IU features performance for recognizing people with none or minimal depression severity from people with mild or higher depression severity is 0.56, which improves to 0.60 when combined with sociodemographic features. Lastly, the IU features performance for recognizing people presenting suicide risk symptoms is 0.54, which improves to 0.57 when combined with sociodemographic features. In all cases, the sociodemographic features alone achieve the best accuracy, ranging from 0.59 to 0.73.

To uncover existing associations between internet usage and depression or suicide risk, this study uses hierarchical mixed-effect models with study participants as random effect to account for individual-level characteristics. The regression analysis reveals that the daily count of application views, the count of application views during the night, the total time spent on chat and messaging platforms, the time spent on message boards and forums, and the number of job-related URLs all have statistically significant positive associations with depression. For suicide risk, it is found that the time spent on chat and messaging platforms, the number of health-related applications, and the number of job-related URLs have positive statistically significant associations with suicide risk severity. Collectively, the results advocate for a comprehensive and inclusive approach to mental health assessment that integrates both traditional sociodemographic factors and emerging internet usage patterns.

Keywords depression, suicide risk, internet usage, browsing behaviour, digital phenotype, web browsing, app usage, desktop, mobile

Preface

I want to thank Professor Juhi Kulshrestha and Professor Talayeh Aledavood for their guidance and expertise.

I also want to thank the members of the Aalto Digital Traces group for their kind support.

Helsinki, 29 January 2024

Emilia Marchese

Contents

Symbols and abbreviations

Symbols

β	hierarchical mixed effect model fixed effect coefficient
β_{std}	standardized hierarchical mixed effect model fixed effect coefficient
χ^2	Chi-square from chi-square distribution

Operators

$\sum_i$	sum over index i

Abbreviations

IU	Internet Usage
GDP	Gross Domestic Product
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ICD-11	International Classification of Diseases, Eleventh Edition
BA	Balanced Accuracy
CI	Confidence Interval
PHQ-9	Patient Health Questionnaire, nine question version
PHQ-9-Q9	Patient Health Questionnaire, question 9 on suicide ideation
RF	Random Forest
XGB	XGBoost Classifier
LR	Logistic Regression
SVM	Support Vector Machine
RBF	Radial Basis Function
HMM	Hierarchical Mixed Effect Model
RFECV	Recursive Feature Elimination with Cross Validation
VIF	Variance Inflation Factor
ICC	Intraclass correlation coefficient
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

1 Introduction

Depression is a leading cause of disability worldwide [1], affecting about 5% of individuals globally [2]. Despite its high prevalence, it is estimated that 50% of the people suffering from depression are not recognized or adequately treated [3]. During the clinical assessment, clinicians often rely on self-reported questionnaire data to assess mental health conditions [4], through surveys that are usually administered sporadically and might spur individuals to provide socially desirable answers [5]. Digital phenotyping is the moment-by-moment quantification of one individual's state through digital devices [6], and can overcome some of the limitations of survey-based mental health assessment. More importantly, it has the potential to be a large-scale early detection tool to recognize people with depressive symptoms or at risk of suicide.

Previous studies have leveraged data from digital devices, such as actigraphy and sensor data from smart devices [7], to sucessfully compute clinical characteristics specific to mental states. Our growing reliance on the internet makes it reasonable to consider internet usage patterns as a possible detection tool for mental health conditions. Additionally, the evolving nature of our online interactions raises the question of potential associations between internet use and mental health. At the same time, recent years have seen a growing awareness and concern among people regarding privacy issues, particularly in the context of digital technologies and online activities [8], raising the need of limiting privacy intrusiveness in data assisted health tools to encourage adherence and use.

Few studies have focused on direct measurements of internet usage for depression classification [9][10][11], showing promising results in using internet usage data for depression assessment. The main limitations of these studies are small and homogeneous populations and poor data quality. This thesis aims to fill in the gaps on the potential of internet usage data from desktop and mobile devices for mental health assessment. Using continuous data from mobile and desktop devices from large and heterogeneous populations of about 900 participants per device type, it aims to assess the potential of URL and app usage traces for depression and sucide risk classification by exploring features sets with different degrees of privacy intrusiveness. Additionally, it aims to identify potential associations between internet usage features with depression and suicide risk when controlling for individual level characteristics and sociodemographic factors. It relies on monthly depression and suicide risk assessments using the Patient Health Questionnaire (PHQ-9) [12]. The PHQ-9 scores are used to measure depression severity and the PHQ-9 question 9 (PHQ-9-Q9) scores are used to measure suicide risk severity.

The findings of this study carry implications for advancing our comprehension of utilizing internet usage data in the assessment of mental health. The findings may contribute to the development of more effective and targeted interventions for individuals at risk of depression or suicide, paving the way for personalized approaches in mental health care. The study addresses the following three objectives:

1. **Objective 1**: To quantify **internet usage** (IU) from desktop and mobile traces in terms of volume, temporal and semantic features which can be useful to infer

user behaviours.

2. **Objective 2**: To explore the potential of the **internet usage** feature sets for **depression classification** and **suicide risk detection** with Machine Learning models and **identify the best performing feature set**.

   This objective is addressed by answering the following research questions:

   - ***Q1:*** *How does the performance differ across device type (desktop and mobile)? Which data from which device is more insightful for depression classification?*
   - ***Q2:*** *How does the performance differ for each of the created internet usage feature subsets? Does more privacy intrusiveness relate to better performance?*
   - ***Q3:*** *How does the performance differ between using IU features only (online features), demographic or sociodemographic features only (offline features), and internet usage plus demographic or sociodemographic features (online + offline features)? Is there an improvement in results obtained by including internet usage features compared to the performance achievable with demographic or sociodemographic information?*
   - ***Q4:*** *How does the performance differ for classifying people with none or minimal depression severity (PHQ-9 $<$ 5) from people with mild or greater depression severity (PHQ-9 $\geq$ 5) versus classifying people with no depression symptoms (PHQ-9 = 0) from people with moderataly severe or higher depression severity (PHQ-9 $\geq$ 15)? Can this technology be useful in early depression diagnosis?*
   - ***Q5:*** *What is the performance for classifying people with no suicide risk (PHQ-9-Q9 = 0) from people with suicide risk (PHQ-9-Q9 > 0)? Can this technology be used in early suicide risk diagnosis?*
   - ***Q6:*** *What are the selected features for the IU set which returns the best performance? Which internet behaviours are the most useful in depression classification?*
   - ***Q7:*** *How do the results compare to those achieved in similar studies, when using similar feature sets?*

3. **Objective 3**: To identify which internet use measures **correlate with depression or suicide risk** when controlling for individual level characteristics and sociodemographic factors.

To address the first objective, the URL traces and app traces from mobile and desktop devices from the PHQ-9 period are labelled with domain and app related categories, and pre-processed into time-series of different granularity (URL, apps, sub-level-domains, sub-categories, on-off events, and others) which are representative of different degrees of data coarseness and user behaviours. The pre-processing tackles known data limitations, including inconsistent categorization across data sources,

time-outs on the duration of individual URL views, and the presence of duplicated views. The created time series are used to engineer features identifying the total volume of internet usage (Aggregate Volume), the volume of internet usage by time of day and time of week (Temporal), the volume of internet usage by viewed content (Semantic), the volume of internet usage by viewed content in a specific time period (Semantic Temporal), and the randomness in user behaviours (Entropies and Kullback-Leiber divergences). The creation of the feature sets is motivated by existing psychological studies and known internet usage associations with depression from similar studies. Lastly, the features sets are ranked by least privacy intrusive (Aggregate Volume) to most privacy intrusive (Temporal Semantic, Entropies and KL) on the basis on the information required to create them.

To address the second objective, the study explores the potential of internet usage (IU) features for binary depression classification and suicide risk detection. Using different feature sets, including IU and sociodemographic data, two PHQ-9 binary splits for depression severity assessment and one PHQ-9-Q9 binary split for suicide risk detection are examined. For depression assessment, the best IU performance for desktop devices ranges from 0.54 to 0.61 depending on the split, and for mobile devices, it ranges from 0.52 to 0.59. The best IU performance is often achieved with the aggregate volume features, emphasizing that more privacy intrusiveness does not always relate to better performance. Semantic and temporal semantic sets also show some potential in the classification. Combining IU with sociodemographic data improves accuracy (up to 0.66 for desktop and 0.65 for mobile) but does not surpass the accuracy achieved with sociodemographic features alone (up to 0.72 for desktop and 0.73 for mobile). For suicide risk assessment, the best IU performances is 0.54 for desktop devices and 0.52 for mobile devices, which improves to 0.57 and 0.54 with the addition of sociodemographic data, but does not outperform the accuracies of 0.63 and 0.59 achieved with sociodemographic features alone. It is found that sociodemographic features exhibit the highest potential, emphasizing the need for a nuanced exploration within diverse sub-groups based on age, income, substance use, and gender.

Lastly, to address the third objective, hierarchical mixed effects effect models with panelist random effects are used to find statistically significant associations between the internet usage features and depression PHQ-9 scores and suicide risk PHQ-9-Q9 scores while accounting for individual level characteristics. Sociodemographic features and monthly fixed effects are included to control for possible confounding effects. The hierarchical model analysis reveals that there are statistically significant associations between internet usage features and depression and suicide risk. The analysis on depression reveals that the daily count of app views, the count of app views in the night, the total time spent on chat and messaging platforms, the time spent of message boards and forums and the number of job-related URLs all have statistically significant positive effects on depression PHQ-9 scores. The analysis on suicide risk reveals that the time spent on chat and messaging platforms, the number of health related apps and the number of job-related URLs have a positive statistically significant association with suicide risk PHQ-9-Q9 scores.

This thesis begins with a detailed background in section 2 on the prevalence of

depression and previous research on digital phenotyping for depression diagnosis. The data collected from the WebWell study is presented in section 3.1.1, followed by a description of the extensive pre-processing of the internet usage traces in section 3.2, the feature engineering in section 3.3, and a correlation analysis in section 3.4. The classification frameworks are introduced in section 3.5 and the hierarchical mixed effect models analysis is introduced in section 3.6. The classification analysis aims to assess the potential of the created feature sets on depression and suicide risk detection by exploring different feature selection methods. The hierarchical model analysis aims to find existing associations between internet usage and depression or suicide risk. Lastly, the results from the classification and hierarchical mixed effect models are presented in section 4 and discussed in detail in section 5.

2 Background

2.1 Depression as a mental health condition

This chapter provides an overview of depression as the most common mental health disorder worldwide. Section 2.1.1 reports the global depression prevalence and in the EU, with a focus on prevalence in Germany as the country of interest for this study. It also highlights the main societal costs associated with depression and the serious issue of depression under-assessment in health care. Section 2.1.2 presents the main diagnostic features of depression. Lastly, section 2.1.3 summarizes the known sociodemographic factors and behavioural habits associated with depressive disorders.

2.1.1 Prevalence of depression, costs and underassessment

Depression is the most common mental illness worldwide, with prevalence rates increasing over time in most developed countries [1][13]. Historical data suggest that the number of people with depression worldwide has increased from 172 million in 1990 to 258 million in 2017, representing an increase of 49.86% [1]. It is estimated that depression affects 5% of individuals globally [2], with marked differences across groups. In 2019, 7% EU citizens reported to suffer from chronic depression, at a rate 0.3% higher than in 2014 [14]. In the same year, Germany documented the EU's third-highest depression rate at 11.6%. Additionally, it recorded the second-highest percentage of men reporting depression (9.9%) and the third-highest percentage of women reporting depression (13.1%) in the EU [14]. Recent years have seen a growth in depression prevalence in Europe, with rates increasing twice in the year following the start of the COVID-19 pandemic for most EU countries, especially among young adults [15]. It is estimated that the COVID-19 pandemic has caused a global rise in depression incidence of more than 25% [15]. In view of these trends, the World Health Organization predicts that depression incidence will only grow in the next decade, making depression the leading cause of illness by 2030 [1].

Depression carries a huge financial burden on the healtcare system, affected invididuals, families and the economy as a whole. There are direct costs to the

healthcare and social systems and indirect cost in the labor market. Health care cost include treatment plans for depression, which usually include medication, psychoterapy and clinical follow-up appraisals [16]. Social costs are in the form of social benefits. Indirect costs to the labour market are increased days of sick leave, decreased work productivity, and early retirement. In fact, depression accounts for up to 50% of chronic sick leaves in the EU, and workers experiencing mental health conditions are estimated to be 6% less productive than usual [13]. In Germany, the excess cost for individual suffering from depression is two times higher for direct and 2.2 times higher for indirect excess cost compared to individuals without depression. In 2015, direct and indirect costs related to mental health constituted almost 5% of the German GDP (Gross Domestic Product) and 4% of the EU GDP (600 billion euros), although the number is possibly higher today [13].

Depression also bears significant costs to the affected individuals. Financial costs in the form of medication and therapy sessions are not covered in every country. While the majority of EU nations include psychological treatments within their healthcare systems, the individual expenses associated with early retirement and sick leaves remain substantial. Depression is also the primary risk of suicidal ideation, and 30% of patients who do not respond to two or more antidepressant treatments will attempt suicide at least once. In 2019, 1.3% of all deaths in the EU were suicide deaths, approximately 120 000 people [13]. Moreover, depression is directly associated with a higher risk of developing several other chronic diseases, including cancer, diabetes and cardiovascular diseases [17].

It is clear that depression is a societal issue, not just an health care issue, which emphasises the urgent need to reduce depression incidence. However, it is estimated that 50% of people suffering from depressive disorders are not recognized or adequately treated [3]. The main challenge with depression diagnosis is the lack of trained professionals, stigma surrounding mental health, and misdiagnosis. Misdiagnosis is a significant concern because depression might become chronic for people who are not diagnosed in time, which implies further economic and personal burdens. Additionally, the scarcity of trained professional is estimated to be 200 times higher in low income countries [18], where the percentage of undiagnosed or misdiagnosed people might be much higher. Stigmatized individuals are more likely to conceal symptoms, delay seeking care and when they do, report more physical complaints instead and be less adherent to treatment [13]. This highlights the need to put more emphasis on prevention in addition to treatment, and actively educate the population on mental health to remove the stigmatization on psychological well-being.

2.1.2 Depressive symptoms

The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [17] is a comprehensive classification and diagnostic tool widely used in the field of mental health. It provides standardized criteria for the classification of mental disorders, aiding clinicians in accurate diagnosis, treatment planning, and communication about mental health conditions. The DMS-5 defines major depressive disorder as a condition characterized by discrete episodes of at least 2 weeks' duration involving clear-cut changes

in affect, cognition, and neurovegetative functions and inter-episode remissions [17]. Depression is accompanied by clear changes in behaviour and attitude that may result in serious functional consequences, such as a higher risk of physical illnesses and deterred social and role functioning.

The main diagnostic features for depression identified by the DMS-5 are:

1. **Mood disturbances**: feeling sad, empty, hopeless, discouraged, appearing tearful, or having increase irritability, brooding, ruminating obsessively and worrying excessively over physical health.

2. **Sleep disturbances**: including difficulty sleeping, sleeping too much or sleeping too little.

3. **Loss of interest in doing things**: less interest in hobbies and not feeling any enjoyment in activities that were previously considered pleasurable.

4. **Changes in appetite**: a significant increase in appetite or a significant decrease in appetite.

5. **Changes in psychomotor activites**: including agitation (inability to be still) or retardation (slowed speech and thinking).

6. **Fatigue and loss of energy**: substantial effort required for the smallest tasks, and efficiency with which the task are accomplised may be reduced.

7. **Increased feeling of worthlessness**: sense of worthlessness and guilt of oneself, rumination over past failings.

8. **Impaired ability to think and make decisions**: easily distracted and complain of memory difficulties.

9. **Thoughts of death and suicidal ideation**: may range from a wish to not wake up in the morning to having a specific suicide plan.

Depression may be caused by combination of different prognostic factors. Environmental factors such as high levels of stress, trauma and adverse childhood experiences have been associated with depression. Genetic, temperamental and psychological factors, including a first degree family members with depression and neuroticism have also been known to predetermine depression. Lastly, depression sometimes arises as a result of course modifiers, including substance use, chronic and disabling medical conditions, diabetes, morbid obesity and cadiovascular diseases[17].

2.1.3 Known sociodemographical association

The incidence of depression varies significantly accross sociodemographic groups. On average, females experience 1.5 to 3-fold higher rates of depression than males beginning in early adolescence [17]. Incidence for young adults of age 18-29 is three times higher than the incidence in individuals aged 60 years or older [17]. Both cultural

factors and geolocation contribute to differences in the prevalence of depression, as well as word-wide events such as the COVID-19 pandemic. A recent study on prevalence in adolescence wordwide found that female adolescents and adolescents from Middle East, Africa, and Asia have the highest risk of developing depression [19]. Moreover, low-income groups and people with lower educational attainment are twice as likely to report chronic depression [13].

Behavioral patterns also play a significant role in influencing the development of depression, as supported by empirical evidence in research. People with sedentary lifestyles and a diet with a low fiber intake report higher levels of depression [20]. Substance use [17], including smoking frequency [20], has also been linked to higher depression severity.

2.2 Survey based depression assessment tools

During the clinical interview with the patients, clinicians usually assess depression by using the DSM-5 or the ICD-11 (International Classification of Diseases, Eleventh Edition). The ICD-11 [21], similarly to the DSM-5, is a globally utilized system for categorizing and coding diseases, injuries, and health conditions for statistical and billing purposes in healthcare. In addition, clinicians generally ask the patient to answer a questionnaire to quantitatively assess the presence of depressive symptoms and their severity. Some common depression assessment scales are the Patient Health Questionnaire (PHQ), the Depression Anxienty Stress Scale (DASS) and the Hamilton Depression Scale (HAMD). This study uses the PHQ-9 questionnaire as a self-assessment tool. Section 2.2.1 introduces the PHQ-9 as a depression assessment tool and section 2.2.2 briefly discusses the known limitations with survey based self-assessment in clinical studies.

2.2.1 The Patient Health Questionnaire

The Patient Health Questionnaire comprises 9 items, each aiming to assess one of the nine main diagnostic features for depression presented in section 2.1.2 as defined by the DMS-5. The scale assesses the presence of the symptom in the previous two weeks from 0 (*Not at all*) to 3 (*Nearly every day*). Question 9 of the scale screens for the presence and duration of suicide risk and suicide ideation. The full PHQ-9 questionnaire can be found in Appendix A. The final PHQ-9 score is the sum of the scores for the individual items. The PHQ-9 total score is commonly demarcated with five thresholds of depression severity. These thresholds are reported in Table 1,

2.2.2 Issues with survey assessment

Two of the main issues with survey based self-assessment are false reporting and recall bias [5]. False reporting occurs when the survey taker decides to dishonestly answer the questions. False reporting is often due to stigma related to mental health and societal expectations, which spurs patients to provide socially desirable answers. The incentive to answer dishonestly is lower when the survey is taken anonymously. Recall

Table 1: PHQ-9 depression severity thresholds.

Score	Severity
< 5	None or minimal depression
5 ≤ score <10	Mild depression
10 ≤ score <15	Moderate depression
15 ≤ score < 20	Moderately severe depression
20 ≤ score ≤ 27	Severe depression

bias is a systematic error that occurs when partecipants do not remember previous events accurately and may omit details. Surveys that rely on retrospective information are at risk of recall bias, especially when the time frame of interest is long. Survey responses are also dependent on mood and attitude of the patient on the day the survey is taken. For these reasons, survey assessment are usually only used as a first approach in clinical diagnosis to screen for the presence of certain symptoms, and followed by a more through review by a clinician if possible.

In the context of behavioural studies, a known risk of survey assessment is raising awareness of certain behaviours in the participant. This is known as demand characteristics, where participants form an interpretation of the experiment and change their behaviour accordingly [9]. A way to mitigate this issue is to avoid explicit wording of the specific object that is being studied. A survey on mental health might ask to answer questions about *well-being* instead of *mental health* to avoid the negative stigmatization associated with mental health disorders. It is also suggested to shuffle the questions if the order is not important, and embed additional unrelated items if the time allows.

2.3 Internet use for depression assessment

The increase burden of depression in all countries calls for new measures for depression prevention and early treatment. Moreover, the marked differences in depression prevalence across age, gender, and other sociodemographic factors urges practitioners and researchers to develop more group-specific and culturally relevant intervention programmes.

In the contemporary digital landscape, the pervasive influence of the internet has seamlessly woven itself into our daily existence. Over the past few decades, there has been a surge in our reliance on the internet, transforming fundamental aspects of our lives such as socializing, commerce, and professional endeavors into online pursuits. This escalating integration of our online presence with our offline reality prompts two crucial inquiries, particularly in the context of mental health.

The initial inquiry revolves around the prospect of utilizing our online behavior, now an integral part of our daily routines, as a diagnostic tool for identifying potential mental health disorders. The evolving nature of our online interactions raises the possibility that patterns in digital behavior may offer insights into mental well-being, that may help us fill the holes in mental health diagnosis and prevention. Depression,

in particular, manifests distinct alterations in cognition and neurovegetative functions, making it logical to anticipate the persistence of these behavioral patterns in the online domain.

The second pivotal question pertains to the correlation between internet usage and mental health. As our lives become more entwined with the digital, it becomes imperative to investigate whether the extent and nature of internet use bear any association with mental health. Understanding these potential connections is vital for comprehending the interplay between our digital presence and psychological well-being, particularly in the light of the raising incidence of mental health disorders in our progressively more digitalized world.

To address these questions, section 2.3.1 presents studies that have used self reported measures of internet use to assess associations with mental health. It proceeds by introducing the topic of digital phenotyping as an assessment tool, and summarizes existing literature on the use of internet use data for depression assessment. Section 2.3.2 reports a summary of known association between internet use and depression collected from the analyzed literature.

2.3.1 Internet use associations with depression

Several studies have focused on assessing the association between internet use and mental health by relying on self-reported internet usage frequency.

Hökby S. et al. [22] aimed to assess whether mental effects of internet use were attributable to the content of the internet use or to the perceived consequences of internet use, such as sleep loss and socialization. They recruited 2286 European adolescents and asked them to answer two surveys 4 months apart. The survey included a depression, stress and anxiety assessment (DASS-42), a suicidal tendency assessment (Paykel suicide scale), a problematic internet use assessment (IAT), internet usage questions for 7 different activities (socializing, gaming, school/work, gambling, newsreading/watching, pornography, targeted searches) and perceived consequences of the activities such as finding friends, sleep loss, learning and others. They performed a cross-sectional hierachical regression analysis to predict the DASS-42 total score from the first wave and found that both the time spent on the internet and on various internet activities were statistically significant predictors, but that the perceived consequences of engaging in those activities were more important predictors. Only gaming, gambling and targeted searches had mental health effects that were not fully accounted for by perceived consequences. They additionally performed a longitudinal hierarchical regression analysis to predict changes in overall psychopathology between the two waves using the changes in internet use and perceived consequences as covariates. The longitudinal analysis showed that sleep loss and withdrawal (negative mood) when internet could not be accessed were the only consequences with direct associations with mental health and that perceived positive consequences (e.g. socialization) did not seem to be associated with mental health at all. The study concludes that perceived negative consequences of internet use seem to predict mental health outcomes to a greater extent that the internet activities themselves.

A 3-wave study [23] of duration of three years involving 27507 people in England

aged 50 or older aimed to explore the relationship between internet use and mental health in older adults. The adults were asked to answer questions about their socioeconomic status, life satisfaction (SWLS) and depression (CES-D). Internet use was assessed with a questionnaire on the time spent on communication, entertainment, information access, finances, ecommerce and others. Data was collected through computer-assisted personal interviews, self-completion questionnaires and nurse assessments. They performed a longitudinal analysis using a hierarchical model to observe the effect of internet use on mental health. They found that internet use frequency was not longitudinally associated with depression, but that there was a positive longitudinal effect of using the internet daily compared to monthly or less on life satisfaction. They also observed that sociodemographic factors moderated the association between internet usage frequency and mental health, and the association was the strongest for those with an educational degree. Additionally, they noted that using the internet for communication purpose, specifically email use, was associated with better mental health and that using the internet for information access, specifically job searching, was associated with worse mental health.

While self-reported measurements of internet use have been shown to be sometimes sufficient for assessing associations with depression, they can hardly be used as an assessment tool for depression prevention and diagnosis. Firstly, they are prone to recall bias, and secondly they require active participation from the patients in completing the assessment. A more objective, continuous and passive quantification could address these limitations. Digital phenotyping is the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices [6]. Digital phenotyping relies on continuous passive data gathering from digital devices, and requires no intervention from the observers [7] or the patient, mitigating both the cognitive bias of the clinitians and any self-reporting bias from the patient. Previous research has shown that digital phenotypes can infer the clinical characteristics specific to mental states, and sometimes with better precision that can be achieved by clinitians [7]. This highlights the potential of digital phenotyping as a powerful tool that can address the under-assessment and misdiagnosis gap in the mental health field. Previous reasearch has mostly focused on sensor data from smartphones and wearable devices for mental health prediction, using digital biomarkers to infer known behaviours associated with depression, such as sleep disturbances and levels of physical activity from actigraphy and motor sensors. Fewer studies can be found on digital phenotyping using direct measurements of internet usage data for depression assessment.

Katikalapudi R. et al. [9] made one of the first attempts to monitor internet usage data and relate it to mental health. They used data from the Missouri S&T campus CiscoNetflow network to explore associations between depression and internet usage for college students. They monitored the internet package flows of 216 college students for the duration of 45 days, and asked them to respond to a one-time depression assessment survey (CES-D). They used distribution difference tests to show that students with depressive symptoms had higher average packets per flow, higher remote file octects, higher email usage and higher entropy in the flow duration. They speculate that the higher average packets per flow might be an indication of streaming and gaming, and the the higher entropy of duration an indication of frequent switching

between tasks.

Katikalapudi R. et al. were only able to measure internet usage through packet flows as a proxy. A later study [24] collected internet browsing URL time series with a desktop plugin for 47 chinese undergrands for the duration of 4 weeks and used a support vector machine (SVM) model to predict the classification accuracy across 7 mental health features (somatization, obsessive compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anziety, paranoid ideation, psychoticsm) assessed with a one time mental health survey (SLC-90). The pre-processing of the logs included 53 features, such as the number of accessed URLs, time of access and the duration of accessing social networks. They achieve a good classification accuracy for all mental health outcomes (70-100%).

More recently, Purwandari B. et al. [25] aimed to assess internet addiction and mental health status (normal, depression, anxiety) from web browsing histories collected from 30 indonesian undergrads for the duration of 5 weeks. They assessed their internet addiction level (IAT scale) and their mental health status (GHQ-12) with a one-time survey assessment, They grouped web behaviour into five types: information retrieval, instant-messaging, social networking services, leisure, and online shopping. The extracted feature for each type was the number of accessed URLs divided by the number of accessed days for each undergrad. They compared the classification performance for the IAT status and the mental health status using three different classifiers: SVM with a radial basis function (RBF) kernel, Random Forest, and Gradient Boosting. They achieved the best classification accuracy with the SVM classifier, 66% for IAD status and 65% for the mental health status.

Some studies have focused on depression classification from mobile app usage and internet browsing on mobile. Yue C. et al [10] extracted internet usage characteristics on smartphones by collecting coarse-grained meta data (source and destination IP addresses and corresponding application-level and transport-layer encryption) from 79 chinese university students for 6 months. The participants were prompted to complete the PHQ-9 survey every 14-days via a custom app and their baseline responses were also validated with an initial screening interview with a clinician. Using the mobile meta-data, they extracted usage sessions from the mobile traces for each PHQ-9 interval and identified three categories of internet usage features: volume features in terms of the amount of traffic in bytes, aggregate usage based features in terms of the number of sessions and the total duration in the PHQ-9 interval and during specific periods (morning, afternoon, evening, night) and usage feature by category for email, gaming, shopping, social, video, audio, and study in terms of number of sessions and duration. They used a SVM model with RBF kernel to classify the features extracted from the individual PHQ-9 intervals. They used leave-one-out user cross validation to prevent leakage in training and validation. They achieve F1 scores between 0.6 and 0.7 for iOS users and between 0.56 to 0.8 for Android users, with the aggregation of features (bytes volume + category based total duration and usage sessions + total duration and usage sessions) returning the best F1 score of 0.7 for iOS user and the best F1 of 0.8 for Android users.

A similar study [11] recruited 456 participants from MTurk portal and asked them to report their gender and age, complete a depression assessment questionnaire

(BDI-II) as well as report their mobile usage, number of calls, duration of calls, and app usage in the 14 days prior. The size of the population assessed makes it the largest among the studies which have employed direct data collection presented in this section. The participants were asked to download two apps from which they could observe the usage statistics for the previous 14 days to report in the questionnaire. The participants were classified as having no to minimal depression symptoms, or mild to severe symptoms. Eight different classifiers were trained and tuned on the features: Classification And Regression Trees (CART), Gradient Boosting Machines (GBM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Neural Networks (NN), Random Forest (RF), SVM with linear kernel and SVM with RBF kernel. The classifiers were trained on 10 different train and test split and the results were averaged across splitting seeds. The classifiers were trained using the mobile usage features only, and the mobile usage features in addition to age and gender information to compare the results. The classification with the mobile usage features achieved the highest AUC score of 0.75 with the random forest classifier, and the classification with the mobile usage features and demographic features achieved the highest AUC score of 0.78 again with the random forest classifier. Feature importance analysis revealed that the number of calls made daily was the most important feature in classification for the random forest model, followed by the average daily duration, the number of contacts saved on the device and the time spend on social media. Distributional tests revealed that there are statistically significant differences on several of the app usage features between the two groups.

2.3.2 Identified internet use associations with depression

This section provides a non-comprehensive list of internet use and mobile use features that have been shown in the literature to be associated with depression. The features are summarized in Table 2, with the direction of the association if known.

2.3.3 Main limitations of previous studies

The existing literature on depression assessment with internet use data presents several limitations, which this study aims to address.

The first limitation pertains studies which have used self-reported measurements for internet use data [22][23]. These studies are affected by the drawbacks of survey assessment presented in section 2.2.2, namely the possibility of recall bias and false reporting.

The second limitation is analysis limited to small population sizes and/or to specific sub-groups. This limitation pertains most of the presented literature, which have small population sizes and often populations that are not representative of the general demographic. Numerous studies have concentrated on adolescents [22] and college students [9][25][10] and a few have targeted older adults [23]. Only one of the analyzed studies [11] with direct measurements of usage data includes a diverse population.

The last limitation that this study aims to address is the quality of the internet usage data and the limitation of the analysis on one type of device. Most of data-based

Table 2: A non-comprehensive list of internet usage and mobile usage features that have been associated with depression in the existing literature.

Feature	Source
Email	Email usage as a form of communication was negatively associated with depression in the older population [23] and in studies with college students [10].
Social-networking	People with mild to severe depression had higher time spent on social networking apps [11]
Communication and instant messaging use	In the older population, using the internet for communication purposes was protective of depressive symptoms [23]. Communication with ones social circle only has been sometimes negatively associated with depression [26] while other studies have found instant messaging to be positively related with depressive symptoms [27]
Games	For adolescents, gaming was a significant predictor of mental health [22][25]
Gambling	For adolescents, gambling was a significant predictor of mental health [22]
Targeted searches	For adolescents and the older population, targeted searches were a significant predictor of mental health [22][23]
Shopping	Shopping disorder and online shopping have been associated with depression [28]
Job related	For the older population, the frequency of job related targeted searches were significant predictors of mental health [28]
Vaguebooking on boards and socials	Vaguebooking, the practice of making a post on social media, was predictive of suicidal ideation in adolescents, which is a depression symptom [29]
Number of calls and call duration	Lower number of calls received and initiated, and lower call duration were predictive of depression [11]
Average daily duration	People with mild to severe depression had higher daily mobile usage [11]
Streaming	Higher packets per flow were indicative of streaming and positively correlated with depression for college students [25]

studies have focused on mobile devices [10][11]. The studies on web browsing from desktop devices are very limited [9][24], and often focus on proxy measurements of internet volume [9], from which it is not always possible to identify the content of the activity.

This study addresses these limitations as follows: the internet usage data is collected moment-by-moment from desktop and/or mobile devices. The data collected are

detailed internet browsing time series with complete URL queries and app usage series including app names. Both the URLs and apps are categorized by their content. The population studied is varied and representative of the general German population. It includes a large sample of approximately 900 individuals per device type, making it the largest population size among similar studies. The data is additionally collected both from desktop devices and mobile devices, so that an analysis can be done for internet usage from different sources. This allows for a comparison on how the associations with depression and the potentials for assessment differ across internet usage data from different devices. Additionally, the data is collected for the duration of several months, and psychological well-being is assessed on a monthly basis, allowing to conduct both a cross-sectional and a longitudinal analysis.

The drawbacks that are not addressed in this study are the reliance on self-reported depression symptoms from the PHQ-9 assessment, and that the population, despite being in the largest one assessed this far, is culturally confined to Germany.

3 Methods

3.1 Data

3.1.1 Data collection: WebWell Longitudinal study

The data was collected as part of the WebWell Longitudinal Study. The study aims to assess associations between internet use and psychological well-being. Desktop web browsing data and mobile usage data is collected continuously for a period of six months and panelists are asked to respond to a survey on psychological well-beign on a monthly basis. A third party company runs a population-representative panel who can be invited to participate in different studies. The company collects the data and prompts the panelists to answer the survey via a mobile app or website. The monthly survey includes the PHQ-9 questionnaire to assess the depression severity of the respondents. The baseline survey was launched on July 2023 and 1490 panelists responded to the survey. Of these, 1066 provided desktop traces, 978 provided mobile traces, and 554 gave access to both mobile and desktop traces. In addition to validated scales to assess psychological well-being, the baseline survey included questions on demographics, income, education level, urbanization and substance use.

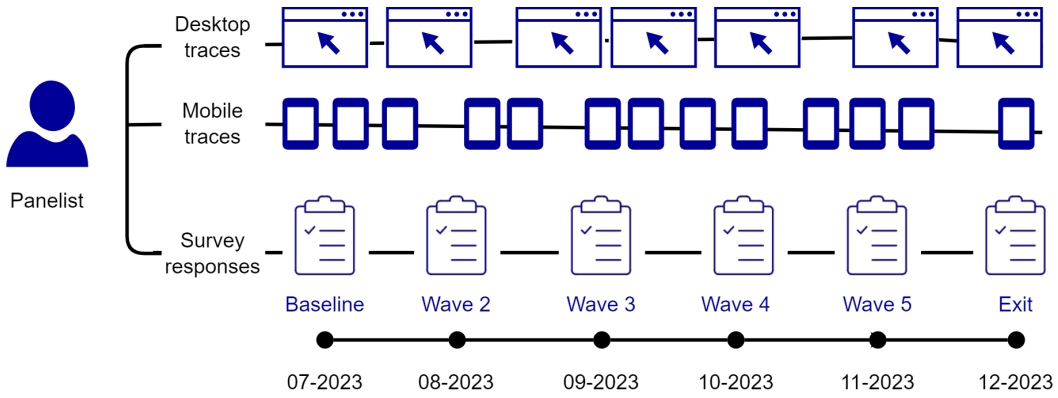


Figure 1: Overview of WebWell Longitudinal study: surveys to assess psychological well-being are administed once a month for six months. Mobile traces and/or desktop traces are collected continuously for each panelist for the duration of the study and for several months preceding.

For the purpose of this thesis, data from different devices is not joined even if a panelist provides both desktop and mobile data. The correlation, classification and regression analysis are carried-out independently for each device type.

The cross-sectional correlation analysis presented in section 3.4 and the cross-sectional classification analysis in section 3.5 are performed using the data from the baseline survey and the associated PHQ-9 interval, which includes the traces from two weeks prior the survey. The longitudinal hierarchical mixed effect models regression analysis presented in section 3.6 uses the data collected from the first three survey waves and their associated PHQ-9 intervals.

3.1.2 Mobile and desktop traces

The data is collected from two types of devices: **desktop** devices and **mobile** devices. A desktop device may be a laptop or a desktop. A mobile device may be a phone or a tablet.

Desktop traces are URL views on a desktop device. Mobile traces are URL views and app views on a mobile device.

For desktop devices, the data is collected via a browser plug-in. The plug-in tracks the active tab of the browser and saves its URL as a new observation in the panelist traces, with the duration as the time spent on the active tab. There is a 180 second time-out if no mouse movement is detected on the active tab. For mobile devices, the URL visits and app visits are collected using a proxy for all requests made by the device, with the duration as the time spent on the activity (URL view or app view) when it is in the foreground and not idle. A timeout of 125 seconds is applied after which the device is considered inactive in the absence of interactions with the touch screen.

Figure 2 and Figure 3 show a simplified example of a sample of desktop and mobile traces after merging with the provided category files. The category files include domain categories from the domains of the URL views, which are provided by a third-party categorization tool called the Webshrinker API [30]. These categories are merged with the URL views on the domain name. The category files also include app categories from the PlayStore to be merged on the app ID for app views. It is possible, that a domain or app doesn't have a category, in which case it is not dropped from the panelist's traces but instead labelled with category *uncategorized*. More details on the categorization are presented in section 3.2.1.

The panelist ID `pid` uniquely identifies each panelist. For desktop devices, each URL view appears as a new row in the traces of the panelist. The duration field indicates how long the panelist observed the URL before moving to the next one or before a timeout. The mobile traces include both URL views and app views. Each view is represented as a new entry in the panelist time series. If the action is related to an app, the URL field is empty. If the action is related to surfing the web via the browser app, the `app_n` field is empty and the URL is reported in the `URL` field.

pid	url	domain	Webshrinker category	used_at	duration
1	google.com/search?...	google.com	search-engines and portals	10/10/10 10:10:10	490
1	netflix.com/watch?...	netflix.com	entertainment, streaming-media	10/10/10 10:15:10	900

Figure 2: Desktop traces: each different URL visit appears as a new observation in the traces. Webshrinker API categories are provided for the domains. One domain may have more than one Webshrinker category.

pid	url	app_n	domain	Webshrinker category	app category	used_at	duration	connection
1	google.com/search?...	NaN	google.com	search-engines and portals	NaN	10/10/10 10:10:10	490	wifi
1	NaN	Neflix	NaN	NaN	Entertainment	10/10/10 10:15:10	900	cellular

Figure 3: Mobile traces: URL visit and app visits appear as new observations in the traces. App categories are provided from the PlayStore. Webshrinker categories are provided for the domain of the URL visit. The internet connection at the time of the observation (wifi, cellular, unknown) is also given for each view.

3.1.3 Panelist selection

The PHQ-9 questionnaire aims to assess depressive symptoms in the two weeks preceding the time of the survey. The timestamp at which each panelist takes the survey is used as the end of their PHQ-9 interval, and the start of the PHQ-9 interval is calculated as 14 days prior to the survey start-time. For each wave, the panelists are selected such that they have at least one observation in their PHQ-9 interval and one observation before or in the first day of their PHQ-9 interval. Panelists that have identified their gender as *non-binary* in the baseline survey are not included. With these criteria, 894 desktop users and 874 mobile users are selected from the 1490 respondents of the baseline survey for the first PHQ-9 interval. It was decided not to use stricter boundaries, for instance selecting participants having activity in at least half of the days in the PHQ-9 interval, to avoid removing participants who might have lower levels of activity as a result of a depressive episode. The minimum dates condition was set to mitigate the risk of selecting people with just one observation in the PHQ-9 interval who might have accessed their devices only to respond to the questionnaire.

3.1.4 Sociodemographics and PHQ-9 score distributions for selected panelists at baseline

The cumulative distributions for the PHQ-9 scores for the panelists selected from the baseline survey is shown in Figure 4. As can be observed, almost half of the selected panelists for either device type have none to minimal depression severity (PHQ-9 < 5), and about 25% of the participants have moderate or higher depression severity (PHQ ≥ 10). This would imply that the proportion of depressed people in our dataset is significantly higher than the share of people reporting chronic depression in Germany in 2019 [14].

Figure 5a shows the distributions of several sociodemographic variables measured at baseline. The distribution of gender is quite balanced for both device types, while the age of the population is slightly biased towards people having 40-60 years. This reflects the age distribution in the German demographic, with people aged 40-59 making up the largest age group in the German population [31].

Figure 5b shows the distribution of PHQ-9 scores by sociodemographic variables. PHQ-9 scores distributions are consistent with the existing literature: women have on

average higher PHQ-9 scores than men and mean PHQ-9 scores decrease with age, with people aged 18-29 having the highest average scores in our population. Mean scores decrease with the education level and monthly income, while there is no clear association that can be observed from the figures for political orientation, urbanization, and tobacco use frequency. With the exception of the slightly higher PHQ-9 scores, the distributions of the population seems to mirror that of the overall demographic.

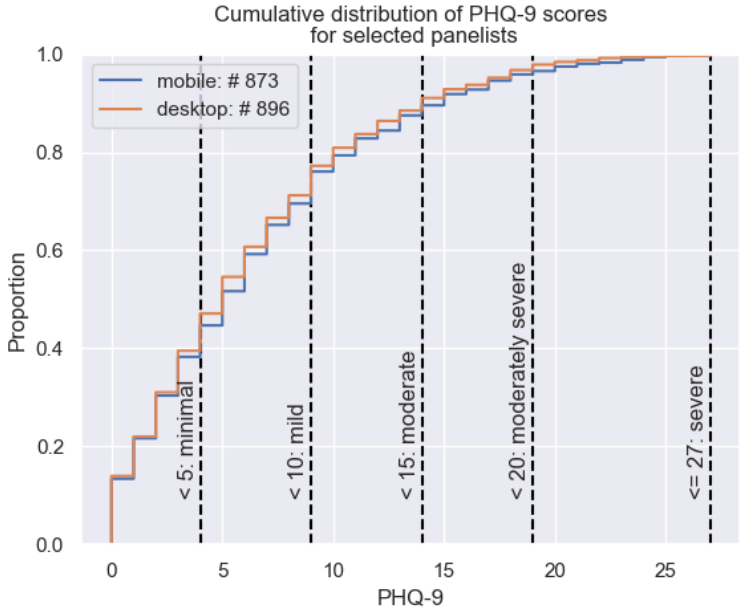
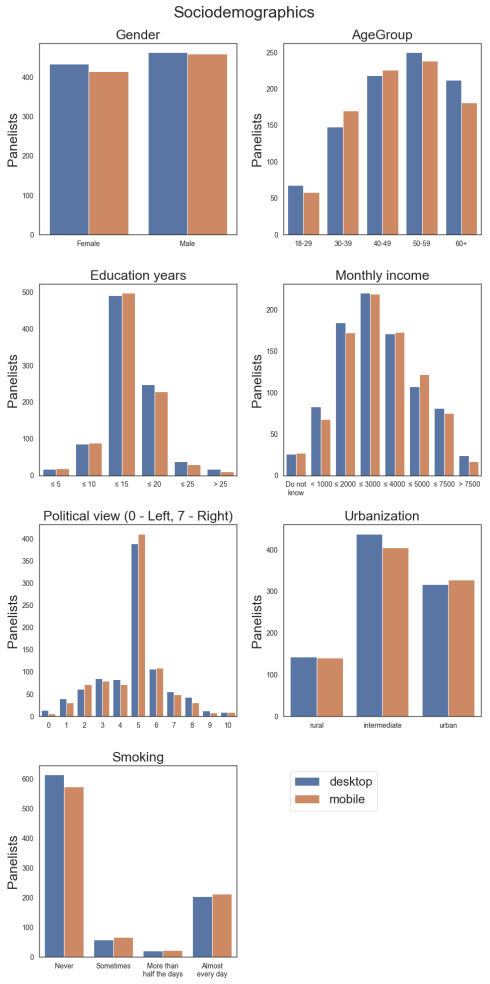
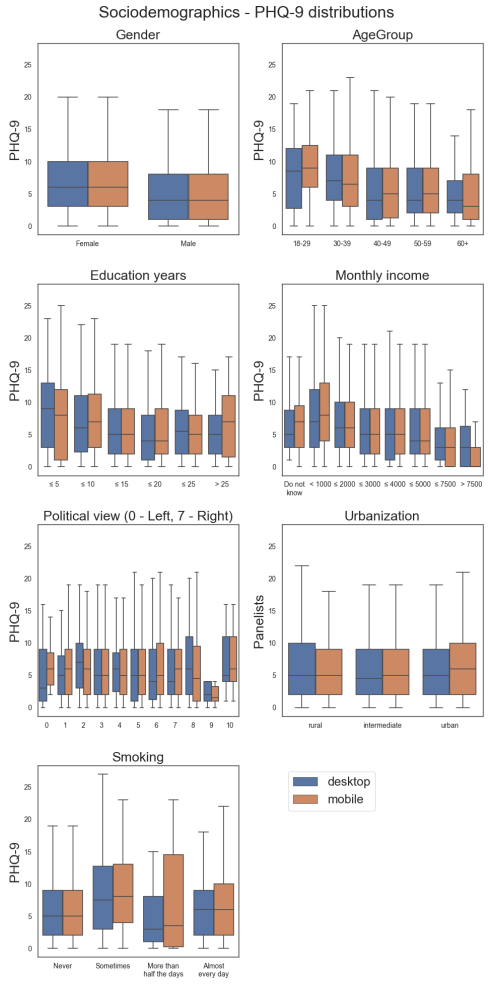


Figure 4: Cumulative distribution of PHQ-9 scores for selected panelists at baseline



(a) Distributional plots for the sociodemographics variables.



(b) PHQ-9 distributions by sociodemographic variable.

Figure 5: Distribution plots for sociodemographic variables (left) and PHQ-9 distribution plots by sociodemographic variable (right).

3.2 Pre-processing

The desktop and mobile traces for the PHQ-9 interval are enriched with a more detailed categorization and preprocessed to take into account the known limitations of the collection framework. This is explained in deail in section 3.2.1. Section 3.2.2 describes how app views are categorized with the domain sub-categories to have a comparable categorization across URL views and app views. The sub-categories are then aggregated into parent categories and interactivity categories to depict different user behaviours. Section 3.2.5 describes how desktop and mobile traces are processed into time series of different granularity that are later used for feature creation as described in section 3.3.

3.2.1 Refined sub-categories

The domains of the URL traces come with 45 Webshrinker categories from the Webshrinker API. A domain may have one or more Webshrinker category. As an example, the domain `netflix.com` is assigned the categories {*entertainment, streaming-media*}. Approximately half of the domains present in the first PHQ-9 interval traces do not have a category, which constitutes about 10% of the URL views.

The Webshrinker categories, while detailed, do not include categories of interest such as email use. Previous research has found relevant associations between email usage and mental well-being [23], highlighting the need to include this feature in this analysis. It is also of interest to observe whether there are potential associations between productivity and work activities online and depression. Therefore, two new custom categories, *email* and *productivity* are created from specific sub level domains and added to the Webshrinker categories. Additionally, the *tools* category is added for the app traces for default apps such as launcher and home apps. The addition of these categories is explained in more detail in Appendix C. Lastly, all domains without categories are labelled as *uncategorized*.

After adding the custom categories *email*, *productivity* and *tools* to the Webshrinker categories, the final set of sub-categories contains the 47 sub-categories shown in Table 3. This enlarged set of categories will be referred to as the sub-categories set in the analyses.

3.2.2 Re-categorization of app views

After redefining the sub-categories set, the app names for the app views are categorized using these sub-categories to have a consistent categorization across URL views and app views. App views come with 12 app categories from the PlayStore, but a brief analysis of these app categories reveals that they are often inaccurate and lack the level of details provided by the final set of sub-categories. The re-categorization into sub-categories is done in three steps. First, the top 600 apps by app view count are collected and categorized as one or more sub-category by string matching the app name to a known categorized domain and manually categorizing the app if no match is found. The final set of top 600 categorized apps is checked for misclassified apps and all mistakes are fixed. The top 600 apps constitute 95% of all app views in the first PHQ-9 interval. Therefore, to re-categorize all app views, the following three-steps approach is employed:

1. If the app name appears in the top 600 categorized apps, use the sub-categories from the categorized set.

2. If the app name does not appear in the top 600 categorized apps, attempt categorizing the app by string-matching on certain keywords. For instance, apps containing strings *dating* are categorized as *dating and personals*. The set of string for matching is created by observing the top apps. More information can be found in Appendix C.

Table 3: Final set of sub-categories used in the analyses.

Sub-categories
business
entertainment
games
survey
health
real-estate
education
news and media
information-tech
shopping
streaming-media
dating and personals
sports
food and recipes
travel
black-list
advertising
parked
adult
job-related
chat and messaging
economy and finance
religion
vehicles
alcohol and tobacco
search-engines and portals
blogs and personals
media-sharing
gambling
message-boards and forums
illegal-content
drugs
content-server
proxy and filter avoidance
social-networking
translators
weapons
abortion
humor
deceptive
malicious
virtual-reality
government
hacking
email
productivity
tools

3. If no match is found with the previous approach, the PlayStore app category is used to infer the sub-category. As an example, apps categorized as *Health&Fitness* from the app store category are categorized under the *health* sub-category. This is explained in details in Appendix D.

4. All remaining apps are labelled as the sub-category *uncategorized*.

To test the accuracy of this approach, the categorization of a random sample of 1000 app views was personally assessed. Of these, 40 app views were deemed wrongly categorized, which represents an error rate of 4%. The categories for these apps were corrected and added in the categorized top 600 apps, so it is expected that the error rate in the final categorization is lower.

3.2.3 Pre-processing of raw traces

The URL traces have two main limitations, which are duplicated views and the presence of the timeouts. Duplicated views may also occur in the app views.

For URL traces, duplicated views are two consecutive URL views with identical URLs at most one second apart. Duplicated URLs should in theory not appear in the data as the data provider aggregates consequent observations to the same URL into one view by default. Several of the features presented in section 3.3 rely on the count of URL views, and these measures would be inflated in the presence of duplicated URLs, which could lead to misrepresenting user behaviour.

The presence of duplicated URLs was tested using a sample of two weeks from the 1st of August to the 15th of August. This subset contained 3714080 URL views for 1639 panelists. Duplicated URLs represented approximately 12% of the entire data. The analysis on duplicated URLs reveals that 55% of these duplicated URLs are the result of the privacy filter `"xxxxx"` used by the collection framework to hide sensitive information, presumably when the panelist has logged in with their credentials. For instance, a typical duplicated URL would be `instagram.com/xxxxx`, which is likely the result of the user scrolling through reels or stories in the Instagram platform which have different URLs endings that are hidden by the privacy filter. The rest of the duplicated URLs are equal full-length URLs, for instance equal links to the same YouTube video. These are the result of inconsistencies in the collection framework beyond the scope of this work. Regardless, the latter group of duplicated URLs needs to be merged into one view with the duration the sum of its composing duplicated URL views. This is needed to avoid overestimating the number of URL views, and report a realistic duration of URL visits for sub-categories and parent categories. The pre-processing steps for dealing with duplicated URL is shown in Figure 6.

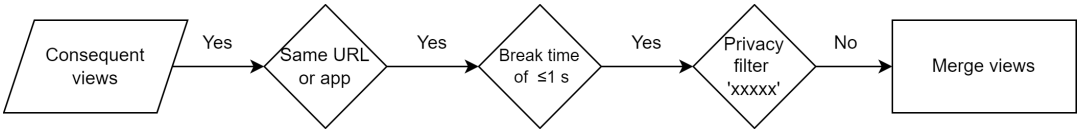


Figure 6: Pre-processing of consequent duplicated URL and app views.

The second limitation is the 180 seconds timeout applied to URL visits. The collection framework applies a URL view inactivity timeout of 180 seconds if no mouse movement is perceived. This needs to be taken into consideration for views with the *entertainment* sub-category that might be recorded as multiples consecutive URL views as a consequence of the timeout.

For instance, consider the scenario where a panelist plays a Netflix episode of the duration of 45 minutes, and pauses the video at the 10 minutes mark for 5 minutes. The collection framework would identify the viewing of this Netflix episode as 4 consecutive URL views with the same URL, the first three of which have active duration of 180 seconds due to the timeout. The first URL view starts when the user starts the video and ends due to the inactivity timeout. The second one starts at the 10 minutes mark, when the panelist touches the mouse to pause the video. This also time-outs. The third one starts at the 15 minutes mark, when the panelist touches the mouse again to restart the video, and the last one at the 45 minutes mark, when the panelist touches the mouse to exit the video. The gaps in between would be considered inactivity time from the collection framework, while it would be more appropriate to consider them as active viewing time and aggregate the four URL views as one having a duration as long as the difference between the end time of the last URL views and the start time of the first timed-out URL view.

A threshold for the gap time needs to be used, under which two consecutive entertainment URL views with the same URL would be aggregated and above which they would be considered as individual observations separated by inactivity time. An analysis of the top entertainment domains reveals that `youtube.com` is the most frequent entertainment domain, followed by other common streaming websites such as `netflix.com` and the German TV platform `zdf.de`. Since the duration of content on TV and series streaming platforms is usually higher than the duration of videos on `youtube.com`, it was decided to use two different thresholds.

For entertainment URL views with domain `youtube.com`, two consecutive URL views with the same URL are aggregated into one observation if the gap time between the start of the first webview and the start of the second webview is less or equal 11.7 minutes. The threshold of 11.7 minutes is the average YouTube video duration in 2018 [32], which is used as a heuristic for the average duration of a YouTube video in this dataset.

For the remaining entertainment domains, a threshold of 45 minutes is used. This is a tentative average of a sample of episode lengths from `netflix.com` and `zdf.de`, as a more accurate average was not found from existing literature. The pre-processing for the timed-out views with entertainment sub-category is summarized in Figure 7. The same pre-processing is applied for app views and URL views in mobile devices.

3.2.4 Re-categorization into parent and interactivity categories

Some of the sub-categories are semantically very similar, for instance *drugs* and *alcohol and tobacco* could be both considered under an umbrella category for substances. It is reasonable to aggregate these categories into parent categories to explore the effect of less granular categories and have each URL and app view identified with fewer parent

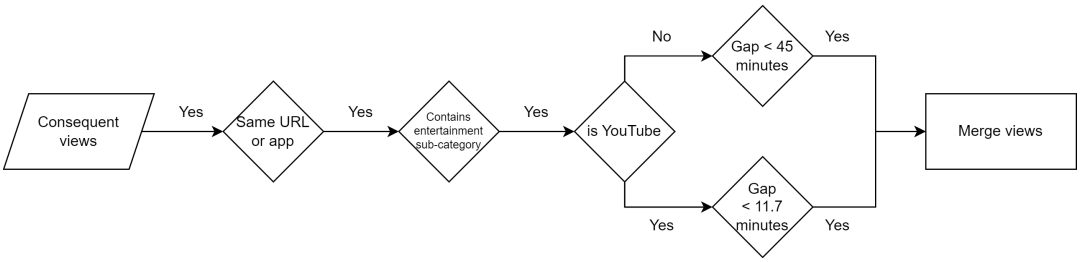


Figure 7: Pre-processing of timed-out view with the entertainment sub-category.

categories instead of multiple sub-categories.

The refined sub-categories are aggregated into 17 parent categories as shown in Table 4 by grouping similar categories together. The choice of where to allocate a sub-category was discussed with two other researchers on the basis of the following three criteria:

- The purpose for which one would observe the sub-category.
- The number of co-occurences with other sub-categories.
- Observing the top domains belonging to this sub-category when in doubt.

For instance, the sub-category *sports* was included under the *Entertainment* parent category after discussing that one of the main reasons to visit this sub-category online is for entertainment (streaming a game as an example) or reading sports news. To decide between the two parent categories *Entertainment* and *News&Media*, it was observed that the sports sub-category co-occurs the most with the *entertainment* sub-category compared to the *news and media* sub-category, and that the top domains labelled as sports were sport streaming services. Therefore, the *sports* sub-category was inserted under the *Entertainment* parent category. Regardless, if a domain has categories {*news and media, entertainment, sports, streaming-media*} it would be labelled with parent categories {*News&Media, Entertainment*} because of the presence of the *news and media* sub-category, which belongs to the *News&Media* parent category. The only exception is the categorization of domains with sub-category *business*. As per the Webshrinker API description [30], *business* is used as a sub-category to a more descriptive Webshrinker category. In fact, the sub-category *business* almost never occurs alone in a domain, and appears to be used by Webshrinker whenever the website belongs to a business, for instance e-commerce websites or a company. Labelling all domains with sub-category *business* with parent category *Business&Finance* would inflate the presence of the *Business&Finance* parent category, making it meaningless when its purpose is to represent business and finance related content. For this reason, it was decided to allocate the *Business&Finance* parent category on the basis of the presence of the sub-category *business* only when *business* is uniquely used to describe a domain. For instance, a domain categorized with sub-categories {*education, business*} will be classified under *Education*, while a domain with the unique sub-category {*business*} will be classified under *Business&Finance*.

Table 4: Breakdown of parent categories.

Parent category	**Sub-categories**
Email&Productivity	*email*, *productivity*, *translators*, *content-server*
News&Media	*news and media*, *government*
Heath	*health*, *abortion*
Education	*education*, *blogs and personals*, *religion*, *food and recipes*
Entertainment	*entertainment*, *streaming-media*, *humor*, *sports*, *media-sharing*
Illegal	*illegal-content*, *malicious*, *deceptive*, *black-list*
Games	*games*
Gambling	*gambling*
Substances&Harmful	*drugs*, *alcohol and tobacco*, *weapons*
Technology	*information-tech*, *hacking*, *virtual-reality*
Shopping	*shopping*
Socials	*social-networking*, *dating and personals*, *chat and messaging*, *message boards and forums*
Business&Finance	*business*, *economy and finance*, *job-related*, *real-estate*, *vehicles*, *advertising*, *travel*
Search&Proxy	*search-engines and portals*,*proxy and filter-avoidance*, *parked*
Tools	*tools*
Uncategorized	*uncategorized*
Other	*survey*

Table 5: Breakdown of interactivity categories.

Interactivity category	**Meaning**	**Sub-categories**
Social	This category includes all sub-categories which involve actively socializing or sharing material with other people.	*media-sharing*, *chat and messaging*, *social-networking*, *dating and personals*, *message-boards and forums*, *media-sharing*, *email*.
Personal	This category includes all sub-categories that require the user some degree of interaction and self-involvement with the activity other than reading and passive viewing.	*games*, *gambling*, *shopping*, *productivity*, *survey*
OtherInter	This category includes the category *tools*, which was deemed separate from all other interactivity categories.	*tools*
Passive	This category includes all sub-categories that are neither in Social or Personal or OtherInter. The category aims to represent passive behaviour such as passive viewing and passive reading.	Any other sub-category, the domain *tiktok.com* and the app name TikTok.

In addition to parent categories, interactivity categories (*Social, Personal, Passive, OtherInter*) are created from the sub-categories to represent the interactivity status of an individual view. The purpose of these categories is to observe whether the interactivity nature of the activity has some association with depression. Table 5 summarizes the intended meaning of each interactivity category and the constituent sub-categories.

The interactivity categories are assigned in the following order of priority: *Social* > *Personal* > *OtherInter* > *Passive*. That is, if a webview has categories {*games, entertainment*}, it will have an assigned interactivity category *Personal*, because the personal interactive category *games* dominates over the passive category *entertainment*. The only exception are URL views with the domain `tiktok.com` and the corresponding TikTok app, which are labelled as having categories {*media-sharing, social-networking*} but were deemed more accurately described as *Passive* instead of *Social* due to the predominant use of the platform as a video viewing and sharing website.

3.2.5 Granularities from internet usage traces

After pre-processing the raw mobile and desktop traces as described in section 3.2.3, the time series are aggregated into different levels of granularity that will be used for

feature creation. For desktop traces, these are:

- **URL traces**: the URL traces from the raw PHQ-9 interval traces, pre-processed as defined in section 3.2.3 by merging duplicated URLs and dealing with the 180 seconds timeout for entertainment URLs.
- **Sub-level domain (SLD) traces**: sub-level domains are created from the URL and domain of the URL traces. Consequent views less than one second apart with the same sub-level domain are aggregated into one sub-level domain view, with the duration as the sum of the composing URL views, the start time as the start of the first composing URL view and the end time as the end time of the last composing URL view.
- **Sub-category traces**: the sub-level domain traces are split into sub-category views equally across the composing sub-categories, such that each sub-category is an observation with duration as the total duration of the sub-level domain view divided by the number of sub-categories of the view.
- **Parent category traces**: equivalent for the sub-category traces, but for parent categories.
- **Interactive category traces**: equivalent for the sub-category traces, but for interactive categories.
- **On-off event traces**: sub-level domain traces are aggregated into on-events by applying a 30 minutes timeout, after which the user is considered inactive if no observation has occurred. The 30-minute timeout for activity was used in a previous study [10]. On events are blocks of activity and off events are blocks of inactivity in the PHQ-9 interval.

Figure 8 gives a schematic representation of the creation of the different granularities from the pre-processed URL traces for the desktop traces. The same pre-processing is applied to mobile traces, except that the app traces consisting only of app views are added as an additional level of granularity, and that sub-category, parent category, interactivity category and on-off events granularities are created from the joint sub-level domain and app traces. Therefore, for mobile traces, the granularities are URL traces, app traces, SLD traces (from URL traces), SLD and App traces (from concatenating SLD and app traces), sub-category traces, parent category traces, interactivity category traces and on-off events traces. A summary of the created traces for each device type can be seen in Figure 9.

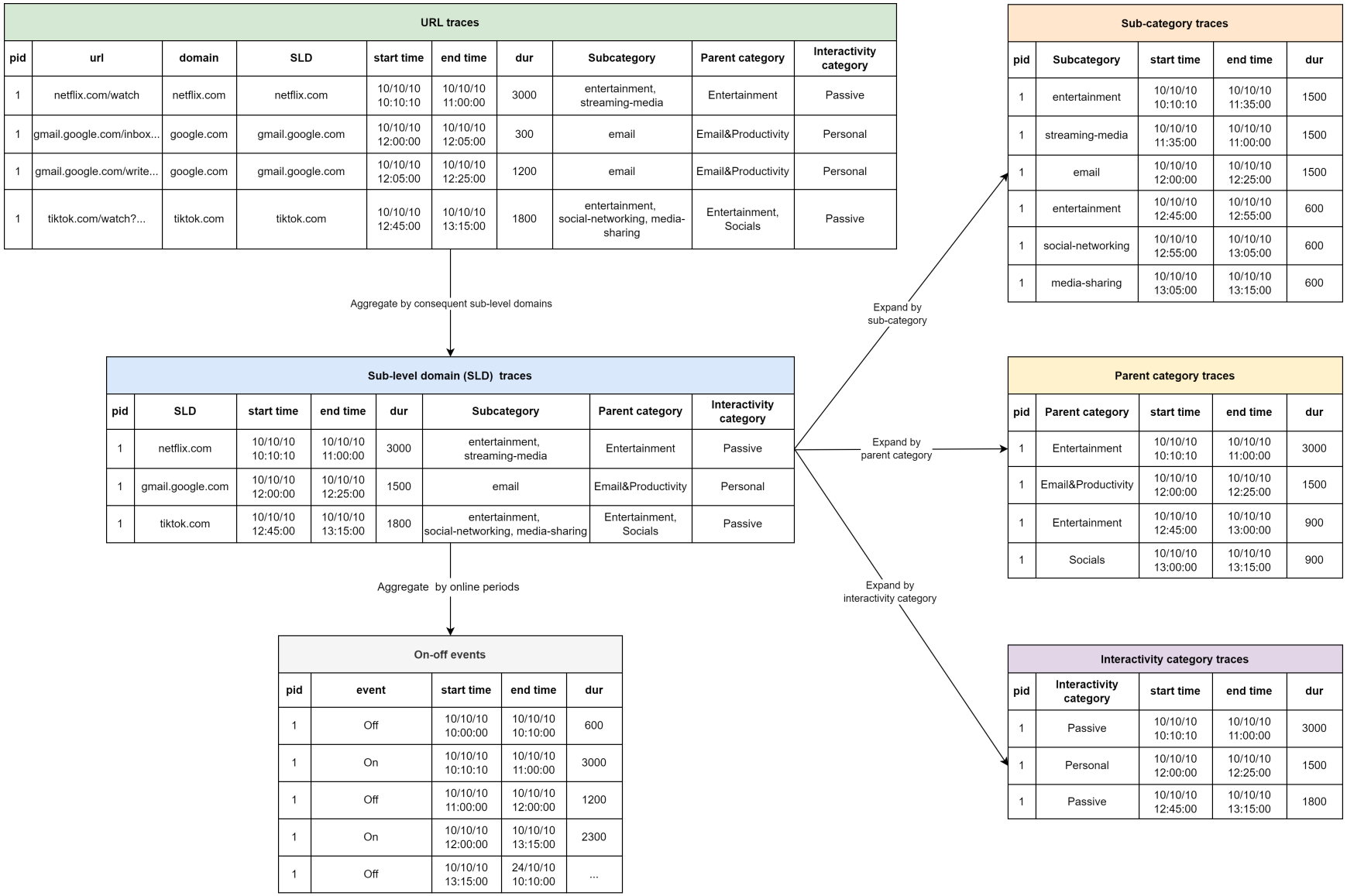


Figure 8: Example of the creation of granularity traces from the URL traces.

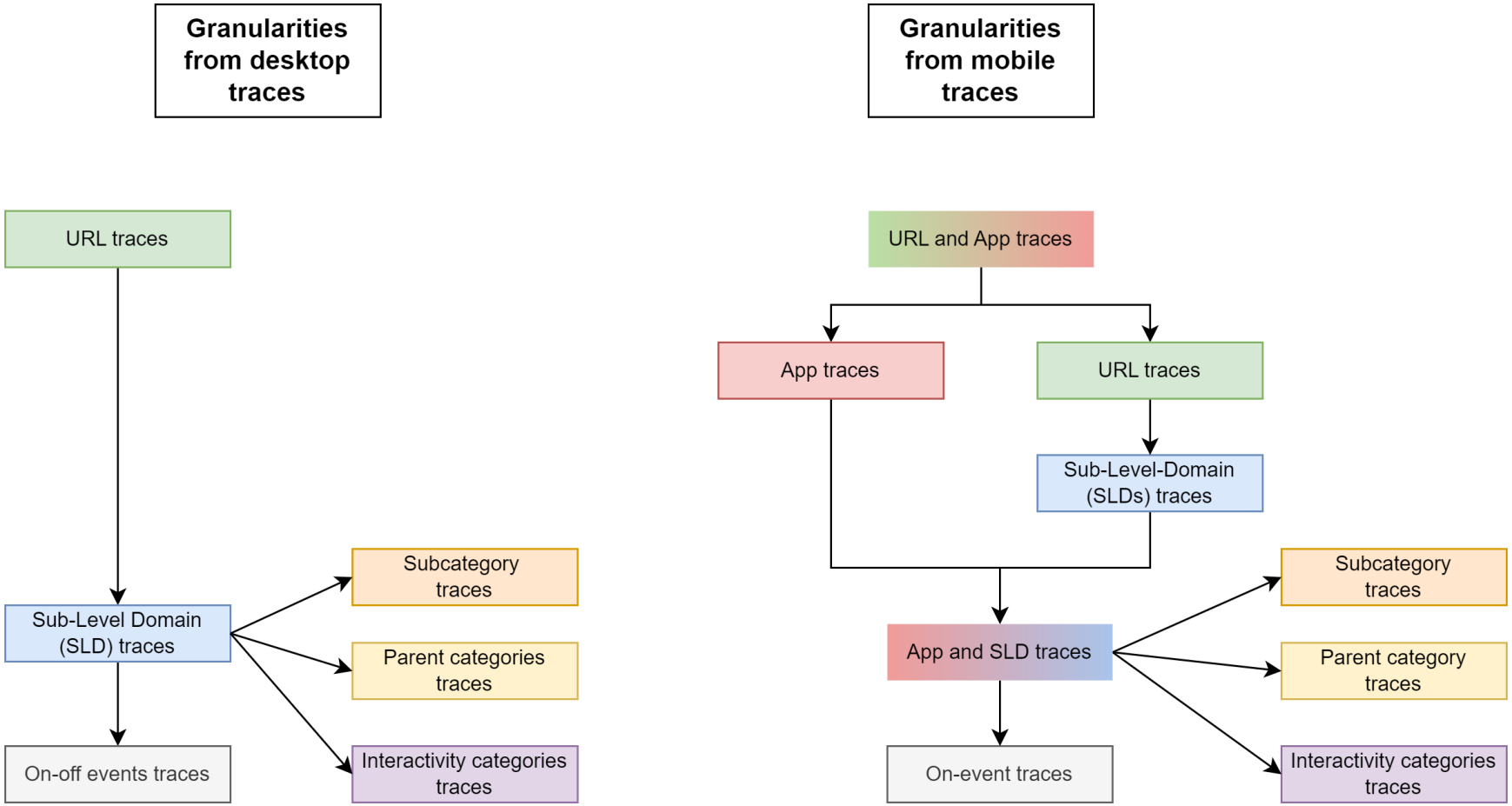


Figure 9: Created granularities for desktop and mobile traces.

3.3 Feature engineering

This work aims to draw conclusions about the associations between web browsing and app usage with depression and its potential for assessment, To achieve this, the traces from the PHQ-9 interval presented in section 3.2.5 are pre-processed to create features indicative of the volume and content of internet use in the two weeks prior the psychological survey assessment. For convenience, app usage and web browsing will be referred to as Internet Usage (IU) from this point forward, even if some app usage activities are not necessarily internet related. The PHQ-9 interval refers to the IU traces collected during the two weeks preceding the survey assessment. Feature engineering is done on the basis of previous studies and aims to infer behaviours that have been associated with depressive symptoms in the psychological literature. Sections 3.3.1–3.3.5 explain in details the feature subsets created from the PHQ-9 interval traces for each device type. The features are summarized in section 3.3.6. Lastly, section 3.3.7 introduces the topic of privacy intrusiveness in relation to the created feature subsets, which motivates part of the workflow of the classification analysis presented in section 3.5 to explore the potential of internet usage data for depression classification.

3.3.1 Aggregate Volume Features

The aggregate volume feature subset includes features that are representative of the total volume and duration of internet usage in the whole PHQ-9 interval. Table 6 summarizes the features in the feature subset and reports the intended meaning of the feature. The device column indicates whether the feature is created for both desktop and mobile devices, or only for mobile devices. Some features are created only for the mobile devices to interpolate quantities that are not provided by the desktop device traces, such as the count of app views and call events.

Table 6: Aggregate Volume feature set. Desktop: 7 features. Mobile: 12 features.

Feature	Description	From traces	Meaning	Device
Ratio of active days in the PHQ-9 interval	The number of dates with at least one observation from the device divided by the total number of dates in the PHQ-9 interval, which is 14.	D: URL M: URL & App	Indicative of the frequency of internet use in the PHQ-9. Depressive episodes may manifest as long periods of inactivity. [17].	D & M
Average daily duration	The sum of the duration of all observations from the panelist PHQ-9 device traces divided by the number of active days in the PHQ-9 interval.	D: URL M: URL & App	Indicative of the total duration of internet use.	D & M
Average daily count of URLs	The count of URL views in the PHQ-9 interval divided by the number of active days.	D: URL M: URL	Indicative of the total volume of web browsing.	D & M
Average daily count on on-events	The count of on-events in the PHQ-9 interval divided by the number of active days	D: On-off M: On-off	Indicative of how many times the panelist was online.	D & M
Average on-event duration	The count of on-event in the PHQ-9 interval divided by the total duration of on-events in the PHQ-9 interval.	D: On-off M: On-off	Indicative of the average length of the online periods of the panelist.	D & M
Average off-event duration	The count of off-events in the PHQ-9 interval divided by the total duration of off-events in the PHQ-9 interval	D: On-off M: On-off	Indicative of the average length of the offline periods of the panelist.	D & M
Average count of URLs per minute	The count of URL views in the PHQ-9 interval divided by the total duration of the URL views in minutes.	D: On-off M: On-off	Indicative of the rate of content consumption. May be useful to infer psychomotor activities (retardation or agitation), which are linked to depression [17].	D & M
Average daily count of apps	The count of app views in the PHQ-9 interval divided by the number of active days.	M: App	Indicative of the total volume of app usage.	M
Average daily time spent browsing on mobile	The total duration of URL views in the PHQ-9 interval divided by the total duration online in the PHQ-9 interval.	M: URL & App	Indicative of the proportion of time spent web browsing on mobile.	M
Proportion of time spent on cellular	The total duration of URL and app views with cellular connection divided by the total duration online in the PHQ-9 interval.	M: URL & App	Proxy measurment of the proportion of time spent outside.	M
Time spent on calls	The total duration of app views with app name "Dialer" , "Phone" or "samsung.incall.iu" in the PHQ-9 interval.	M: App	Indicative of the time spent on voice communication, which has been shown in previous studies to be associated with depression [11].	M
Number of calls	The count of app views with app name "Dialer", "Phone", or "samsung.incall.iu" in the PHQ-9 interval.	M: App	Indicative of the number of calls received or initiated, which has been shown in previous studies to be associated with depression [11].	M

The number of calls and the time spent on calls for the mobile devices is calculated by collecting the app views having app name *Dialer*, *com.samsung.android.incallui* or *Phone*, which were selected as the phone apps of the mobile devices by string searching related strings (*call*, *phone*, *dialer*) and further screening. It's possible that mobile devices of specific brands have different app names and that these measures are under-represented for those devices.

3.3.2 Temporal Features

Past research has shown that temporal features are important in analyzing and understanding human behavior from their digital traces [33]. The temporal feature subset includes features that are representative of the volume and the duration of internet use during specific periods of the PHQ-9 interval. The PHQ-9 interval is divided into time of day and day of the week. The identified time of day periods are the time of days Morning (6–12), Afternoon (12–18), Evening (18–24), Night (24–6). This kind of temporal division has been used in past literature [34]. The identified day of the week periods are Weekday (Monday–Friday), Weekend (Saturday–Sunday). The final set of periods includes the time of day intervals (M-A-E-N), the day of the week intervals (Weekday, Weekend), and combinations of the two, for instance '(Weekday, M)' represents all weekend mornings in the PHQ-9 interval. The final set therefore includes 14 periods. The time of day and day of the week information is added to the granularities using the time and date of the start-time of the observation. For instance, if an URL view starts on Monday 10/10/10 10:10:10, it is labelled as having time of day Morning (M) and day of week Weekday. For each period, the features reported in Table 7 are created.

Table 7: Temporal feature set for period *i* in periods M-A-E-N, (Weekday, Weekend) and combinations. Desktop: 8 features for each period, 112 features. Mobile: 11 features for each period, 154 features.

Feature	Description	From traces	Meaning	Device
Total duration	The sum of the duration of all observations in the period *i*	D: URL M: URL & App	Indicative of the duration in the specific period.	D & M
Count of URLs	The count of URL views in the period *i*	D: URL M: URL	Indicative of the volume of web browsing in the period	D & M
Relative duration	The sum of the duration of all observations in period *i* divided by the sum of the duration of all observations in the PHQ-9 interval	D: URL M: URL & App	Indicative of how much time is spent online in the period in proportion to the total time spent online	D & M
Fraction of URLs	The count of URL views in period *i* divided by the count of URL views in the PHQ-9 interval	D: URL M: URL	Indicative of how much web browsing is done in the period in proportion to the total web browsing.	D & M
Average count of URLs per minute	The count of URL views in period *i* divided by the total duration of URL views in the period	D: URL M: URL	Indicative of the rate of internet consumption in the period. Can be used to infer psychomotor activity, agitation, retardness	D & M
Average off-event duration	The sum of the duration of off-events in period *i* divided by the count of off-events in the period	D: On-off M: On-off	Indicative of the average length of the offline periods of the panelist in the period. Can be used to infer e.g. sleeping patterns.	D & M
Average count of URLs per minute	The count of URL views in period *i* divided by the total duration of the URL views in period *i* in minutes.	D: URL M: URL	Indicative of the rate of content consumption. May be useful to infer psychomotor activities (retardation or agitation)	D & M
Average count of unique SLD per on event	The sum of unique sub-level-domains (SLD) per on event in period *i* divided by the number of on-events in period *i*	M: On-off M: On-off	Indicative of the variability of sub-level-domains viewed in the period whenever the panelist is online	D & M
Count of apps	The count of app views in period *i*	M: App	Indicative of the volume of app usage in the period	M
Fraction of apps	The count of app views in period *i* divided by the total count of app views in the PHQ-9 interval	M: App	Indicative of the volume of app usage in the period in proportion to the total app usage.	M
Average count of unique apps per on event	The sum of unique apps per on event in period *i* divided by the number of on-events in period *i*	M: App	Indicative of the variability of apps viewed in the period whenever the panelist is online	M

3.3.3 Semantic Features

The semantic feature subset includes features that are representative of the volume and the duration of internet use in the PHQ-9 interval for specific categories of

content. The identified categories sets are the sub-categories reported in Table 3, the parent categories reported in Table 4, and the interactivity categories reported in Table 5. For each category of each category set, the features presented in Table 8 are created. In the feature creation, the sub-category *uncategorized* and the parent category *Uncategorized* are omitted because meaningless in this context.

Table 8: Semantic feature set. Mobile: 7 features for each category in each category set. Desktop: 5 features for each category in each category set.

Feature	Description	From traces	Meaning	Device
Total duration	The sum of the duration of all observations with category *i*	D: category M: category	Indicative of the time spent observing category *i*	D & M
Count of URLs	The count of URL views with the category *i*	D: URL M: URL	Indicative of the volume of web browsing for category *i*	D & M
Relative duration	The sum of the duration of all observations with the category *i* divided by the sum of the duration of all other categories	D: URL M: URL & App	Indicative of how much time observing content from category *i* proportion to the total time spent observing content from all categories	D & M
Fraction of on-events with category	The count of on events with category *i* divided by the total count of on-events in the PHQ-9 interval	D: On-off M: On-off	Indicative of whether observing the category is a habitual event	D & M
Average URL duration	The total duration of URL views with the category *i* divided by the total count of URL views with the category *i*	D: URL M: URL	Indicative of the average duration spent on category *i* when browsing.	D & M
Count of apps	The count of app views with category *i*	M: App	Indicative of the volume of app usage with category *i*	M
Average app view duration	The total duration of app views in with category *i* divided by the total count of app views with category *i*	M: App	Indicative of the average duration spent on category *i* on apps	M

3.3.4 Entropies and KL Divergences

This feature subset includes the Shannon entropies and the Kullback-Leibler divergences against random behaviour for the created granularities. Shannon entropies are a measure of the the degree of variability, complexity, disorder, and randomness in the participant behavior states. Previous studies have shown that there might be a positive correlation between the screen-status (on-off periods) normalized entropy and depression [35], highlighting the relevance of including entropy measures in this work. Equation 1 shows the Shannon entropy formula. In the formula, n is the length of the vocabulary, that is the number of possible outcomes. For instance, for on-off events, the possible outcomes are ON and OFF, so $n = 2$. Then x_i is the event $\{x_i : x_i \in (ON, OFF)\}$ and $P(x_i)$ is this probability of the event.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2(P(x_i)) \tag{1}$$

When calculated in base 2 as shown in Eq. 1, the Shannon entropy is measured in terms of bits, and represents the average number of bits needed to encode the information. Low entropy indicates a high level of predictability, certainty and order in the data, while high entropy indicates a high level of unpredictability, uncertainty, or randomness in the data. In addition to the entropy, the Kullback-Leiber divergence, or relative entropy, is calculated against random behaviour. In this context, this may be viewed as another measure of randomness in the behaviour, and gives an indication of how many more bits of information would be required instead if the behaviour of the user were completely random. As panelists may have different vocabularies, for instance the set of apps visited might be different across panelists, this measure gives additional information on the divergence from completely random behaviour. As an example, a panelist may have higher entropy because they have visited more apps, but their behaviour might be more ordered, which would translate as having high KL divergence from random behaviour. Equation 2 shows the formula for the Kullback-Liber diverge of the true distribution $P(x_i)$ against random behaviour $Q(x_i)$. In the formula, $P(x_i)$ is the real probability of the event, while $Q(x_i)$ is the uniform probability of the event $Q(x_i) = 1/n$, which represents the expected probability under completely random behaviour.

$$D_{KL}(Q||P) = \sum_{i=1}^{n} Q(x_i) \log_2 \left(\frac{Q(x_i)}{P(x_i)} \right) \tag{2}$$

For desktop devices, the entropies and KL divergences are created for URLs, SLDs (sub-level domains), sub-categories, parent categories, interactivity categories and on-off events. For mobile devices, the entropies and KL divergences are in addition created for the apps from the app traces and the connection (cellular, wifi, unknown) for URL and app traces. For on-off events granularities, the PHQ-9 interval is split into blocks of 10 minutes, with each block labelled as ON if an ON event occurs in between, or OFF otherwise. This is done because on-off events as described in section 3.2.5 would be an alternating sequence of ON events and OFF events, which would be

meaningless for entropy creation since the probability of each event would be $1/n$. By splitting the PHQ-9 interval into blocks of equal duration, on-off events can be used to measure the randomness of the user behaviour.

3.3.5 Semantic Temporal Features

The semantic temporal features is representative of the duration of internet use for a specific category during a specific period. For each category in the sub-categories, parent categories and interactivity categories sets, the duration during one of the identified 14 periods is calculated. The aim is to explore whether observing a category during a specific time of day is a relevant depression predictor.

Table 9: Semantic temporal features for the M-A-E-N, (Weekday, Weekend) and combination periods for each category in the sub-categories, parent, and interactivity category sets.

Feature	Description	From traces	Meaning	Device
Total duration	The sum of the duration of all observations with category *i* in period *j*	D: category M: category	Indicative of the time spent observing category *i* in the specific period.	D & M

3.3.6 Summary of created features

Table 10 summarizes the features created in each feature subset. The size of the feature set differs by the device type.

Table 10: Summary of the created internet usage feature sets. Periods include M-A-E-N, (Weekday, Weekend) and combinations. Categories include each category in the sub-categories, parent categories and interactivity category sets.

Feature set	**Features**	**Device**
Aggregate Volume	Ratio of active days in PHQ-9	D & M
	Average daily duration	D & M
	Average daily count of URLs	D & M
	Average daily count of on-events	D & M
	Average on-event duration	D & M
	Average off-event duration	D & M
	Average count of URLs per minute	D & M
	Average daily count of apps	D & M
	Average daily time spent browsing on mobile	D & M
	Proportion of time spent on cellular	M
	Time spent on calls	M
	Number of calls	M
Temporal For each period *i*	Total duration	D & M
	Count of URLs	D & M
	Relative duration	D & M
	Fraction of URLs	D & M
	Average count of URLs per minute	D & M
	Average count of unique SLD per on-event	D & M
	Count of apps	M
	Fraction of apps	M
	Average count of unique apps per on-event	M
Semantic For each category *j* in category sets: - sub-categories - parent categories - interactivity categories	Total duration	D & M
	Count of URLs	D & M
	Relative duration	D & M
	Fraction of on-events with category	D & M
	Average URL duration	D & M
	Count of apps	M
	Average app duration	M
Semantic Temporal For each period *i* and each category *j* in category sets; - parent categories - interactivity categories	Total duration	D & M
Entropies and KL For chosen *granularities*	Shannon entropy	D & M
	Kullback-Leiber divergence	D & M

3.3.7 Privacy intrusiveness

With the advent of the digitization and social media, there is a growing awareness and concern among people regarding privacy issues, particularly in the context of digital technologies and online activities. Privacy has also been identified as one of the main features required in data collection in digital phenotyping studies [36]. A recent study on the barriers to engagement on the adoption of mobile apps for depression and anxiety showed that 60% of the participants expressed concerns about the privacy of personal information that is collected by mobile health applications [37]. Additionally, the recent pandemic has brought several governments to adopt increased surveillance measures in the interest of public health, which has further fueled the ongoing debate on finding the right equilibrium between protecting individual privacy rights and addressing collective health concerns [38]. Web browsing traces, specifically, can be easily used to identify individuals. A recent study [39] has shown that knowledge from the four most visited web domains can uniquely identify 95% of individuals, posing a significant threat to digital anonymity in case of data leakage.

Privacy concerns can be a significant factor influencing individuals' decisions not to use digital health technologies. The extent to which privacy concerns impact adoption varies among individuals and is influenced by factors such as awareness, trust, regulatory frameworks, and the perceived benefits of the technology [8]. A relevant part of this study is to explore the potential of different IU feature sets for depression diagnosis, with the intent of identifying useful features but also to understand whether access to more detailed personal information relates to better performance.

Figure 10 ranks the created feature subsets into different levels of privacy intrusiveness on the basis of the information required to create the features. The ultimate purpose is to minimize the amount of data collected and to mitigate the risk of inferring sensitive information in the event of a data leak. In this research, the degree of intrusiveness is therefore measured by the quantity of information required, along with its potential to disclose additional sensitive user details, necessary for constructing the feature set from the internet usage traces. For example, aggregate volume features necessitate knowledge of the duration of individual views. In contrast, temporal features require information on both the duration and date-time of individual views. Semantic features, meanwhile, require details about the duration and category of individual views. However, in this research, semantic features are deemed more privacy intrusive than temporal features due to the category traces potentially revealing sensitive information about the panelist — such as health conditions, religious beliefs and financial status — which cannot be inferred solely from the duration and date-time traces used to construct the temporal features.

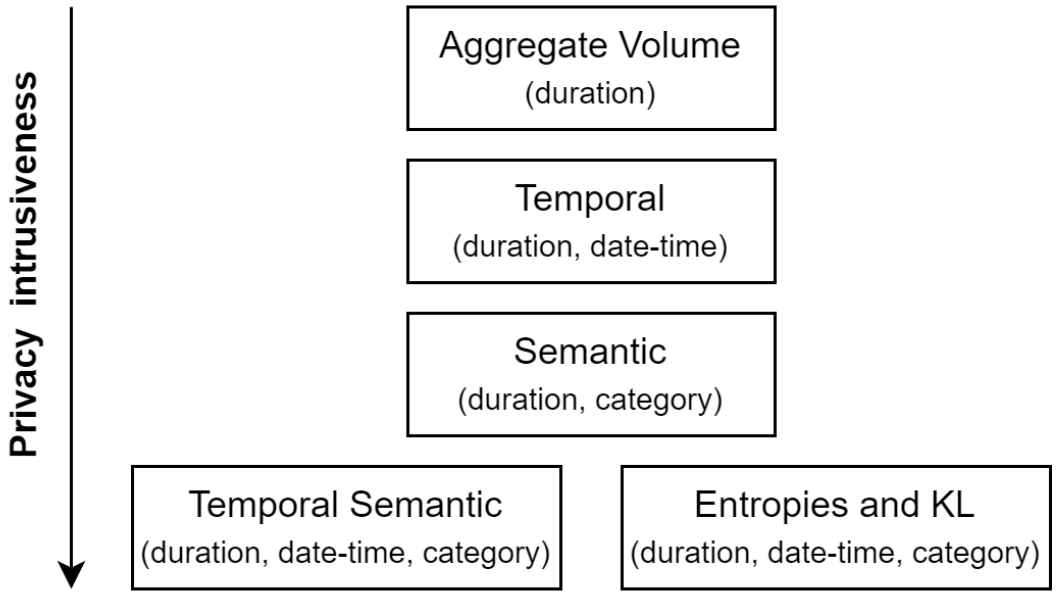


Figure 10: Internet Usage feature sets ranked by privacy intrusiveness in terms of the the quantity of information required, along with its potential to disclose additional sensitive user details, necessary for constructing the feature set from the internet usage traces.

3.4 Correlation Analysis

The aim of the correlation analysis is to explore existing correlations between the internet use features created in section 3.3 and the depression and suicide risk scores. The correlation analysis is performed cross-sectionally only for the baseline PHQ-9 period. The spearman rank correlation between the PHQ-9 score and the PHQ-9-Q9 score and each feature is computed. Rather than applying false discovery rate procedures, such as Benjamin-Hochberg, it is decided to use a more stringent level of statistical significance (0.005) to observe the significant correlations per feature set. The reason this approach is chosen is that different feature subsets have different number of features, and adjusting the p-values family-wise would penalize feature sets with a larger number of features, such as the semantic sets. At the same time, applying the correction on the composite set of all feature sets classifies most features as non significant probably due to the large number of features, which is against the purpose of this correlation analysis. Applying a more stringent boundary on the p-value screens the features without discriminating against larger sets.

The final set of statistically significant features is shown in Table 11 for mobile and in Table 12 for desktop. The correlations are also calculated for the suicide score (PHQ-9 question 9) to observe existing correlations with suicide risk, and reported in Table 13 and in Table 14 for mobile and desktop respectively.

The correlation analysis reveals that existing correlations are weak, and greatest for the sociodemographic variables. For desktop data, morning activity is often negatively associated with depression. For the suicide risk, the statistically significant features are few. Interestingly, the use of *message boards and forums*, such as the Reddit platform, was a significant feature for suicide risk in the mobile devices.

The correlation analysis cannot be trusted to show associations with the PHQ-9 depression score or PHQ-9-Q9 suicide risk score, because the coefficients might be mediated by other variables that are not taken into consideration.

The hierarchical mixed effect models analysis presented in section 3.6 mitigates this limitation by including the effect of sociodemographic variables and other internet use features as covariates. Regardless, the correlation analysis can already give an indication of online behaviours that can be useful in depression diagnosis. The correlation analysis presented here is only for exploratory purposes and not for feature selection. Feature selection is done through more targeted techniques for the classification and the hierarchical mixed effect model analyses.

Table 11: Mobile: PHQ-9 scores correlation results for the mobile device for significant features at significance level 0.005 or lower. **: p value < 0.005, ***: p value ≤ 0.001

Feature subset	Features	ρ
Sociodemographic	age	-0.19***
	gender	0.18***
	income	-0.20***
Aggregate Volume	Average daily duration in active days	0.11**
	Average on-event duration	0.14***
Temporal	('A', 'Weekday'): total duration	0.10**
	('A', 'Weekend'): total duration	0.11***
	('A', None): total duration	0.11***
	('E', 'Weekday'): total duration	0.11**
	('E', 'Weekend'): total duration	0.10**
	('E', None): total duration	0.11**
	('N', 'Weekend'): count of app	0.10**
	(None, 'Weekday'): total duration	0.10**
	(None, 'Weekend'): total duration	0.11**
Semantic Interactive	Social: total duration	0.11**
Semantic Parent	Health: count of URLs	0.10**
	Socials: total duration	0.11***
Semantic Subcategories	health: count of URLs	0.10**
	social-networking: average app duration	0.13***
	social-networking: count of apps	0.14***
	social-networking: fraction of on-events with c...	0.11**
	social-networking: total duration	0.12***
	sports: average app duration	-0.11**
	sports: count of apps	-0.10**
	sports: fraction of on-events with category	-0.10**
	sports: relative duration	-0.10**
	sports: total duration	-0.10**
	streaming-media: average app duration	0.10**
	streaming-media: fraction of on-events with cat...	0.10**
	streaming-media: total duration	0.10**
	tools: relative duration	-0.09**
Semantic Temporal Interactive	('A', 'Weekday') Social: total duration	0.11**
	('A', 'Weekend') Social: total duration	0.11***
	('A', None) Social: total duration	0.11***
	('E', 'Weekday') Passive: total duration	0.10**
	('E', 'Weekday') Social: total duration	0.10**
	('E', 'Weekend') Social: total duration	0.10**
	('E', None) Social: total duration	0.10**
	(None, 'Weekday') Social: total duration	0.10**
	(None, 'Weekend') Social: total duration	0.12***
Semantic Temporal Parent	('A', 'Weekday') Socials: total duration	0.11**
	('A', 'Weekend') Socials: total duration	0.11***
	('A', None) Socials: total duration	0.11***
	('E', 'Weekday') Socials: total duration	0.10**
	('E', 'Weekend') Socials: total duration	0.10**
	('E', None) Socials: total duration	0.11**
	(None, 'Weekday') Socials: total duration	0.11**
	(None, 'Weekend') Socials: total duration	0.11***
Semantic Temporal Subcategories	('A', 'Weekday') social-networking: total duration	0.12***
	('A', 'Weekend') social-networking: total duration	0.11**
	('A', 'Weekend') streaming-media: total duration	0.10**
	('A', None) social-networking: total duration	0.12***
	('A', None) streaming-media: total duration	0.11**
	('E', 'Weekday') social-networking: total duration	0.13***
	('E', 'Weekday') sports: total duration	-0.10**
	('E', None) social-networking: total duration	0.13***
	('M', 'Weekend') social-networking: total duration	0.11**
	('N', 'Weekday') streaming-media: total duration	0.11***
	('N', 'Weekend') search-engines and portals: to...	0.10**
	('N', 'Weekend') streaming-media: total duration	0.10**
	('N', None) social-networking: total duration	0.11**
	('N', None) streaming-media: total duration	0.12***
	(None, 'Weekday') social-networking: total dura...	0.12***
	(None, 'Weekday') streaming-media: total duration	0.10**
	(None, 'Weekend') social-networking: total dura...	0.11**

Table 12: Desktop: PHQ-9 scores correlation results for the desktop device for significant features at significance level 0.005 or lower. **: p value < 0.005, ***: p value ≤ 0.001

Feature subset	Features	ρ
Sociodemographic	age	-0.19***
	gender	0.14***
	income	-0.15***
Temporal	('M', 'Weekday'): count of URLs	-0.10**
	('M', 'Weekday'): fraction of URLs	-0.12***
	('M', 'Weekday'): relative duration	-0.12***
	('M', None): average count of unique SLD per on...	-0.10**
	('M', None): average off-event duration	0.10**
	('M', None): count of URLs	-0.11***
	('M', None): fraction of URLs	-0.14***
	('M', None): relative duration	-0.13***
	('M', None): total duration	-0.10**
Semantic Parent	Adult: average URL duration	-0.12***
	Adult: count of URLs	-0.12***
	Adult: fraction of on-events with category	-0.13***
	Adult: relative duration	-0.13***
	Adult: total duration	-0.12***
	News&Media: average URL duration	-0.10**
Semantic Subcategories	adult: average URL duration	-0.12***
	adult: count of URLs	-0.12***
	adult: fraction of on-events with category	-0.13***
	adult: relative duration	-0.13***
	adult: total duration	-0.12***
	entertainment: fraction of on-events with category	0.10**
	news and media: average URL duration	-0.10**
Semantic Temporal Interactive	('M', 'Weekday') Passive: total duration	-0.10**
	('M', None) Passive: total duration	-0.11**
Semantic Temporal Parent	('A', 'Weekday') Adult: total duration	-0.13***
	('A', None) Adult: total duration	-0.11**
	('M', 'Weekday') Adult: total duration	-0.12***
	('M', 'Weekday') Business&Finance: total duration	-0.12***
	('M', 'Weekend') Adult: total duration	-0.12***
	('M', 'Weekend') Business&Finance: total duration	-0.10**
	('M', None) Adult: total duration	-0.14***
	('M', None) Business&Finance: total duration	-0.14***
	(None, 'Weekday') Adult: total duration	-0.14***
	(None, 'Weekday') News&Media: total duration	-0.10**
	(None, 'Weekend') Adult: total duration	-0.09**
Semantic Temporal Subcategories	('A', 'Weekday') adult: total duration	-0.13***
	('A', 'Weekend') blogs and personal: total dura...	0.13***
	('A', None) adult: total duration	-0.11**
	('M', 'Weekday') adult: total duration	-0.12***
	('M', 'Weekday') business: total duration	-0.10**
	('M', 'Weekend') adult: total duration	-0.12***
	('M', 'Weekend') business: total duration	-0.10**
	('M', 'Weekend') education: total duration	-0.10**
	('M', None) adult: total duration	-0.14***
	('M', None) business: total duration	-0.13***
	('M', None) education: total duration	-0.10**
	('M', None) parked: total duration	-0.11**
	(None, 'Weekday') adult: total duration	-0.14***
	(None, 'Weekday') news and media: total duration	-0.10**
	(None, 'Weekend') adult: total duration	-0.09**
	(None, 'Weekend') blogs and personal: total dur...	0.11**

Table 13: Mobile - Suicide Risk: PHQ-9-Q9 scores correlation results for the mobile device for significant features at significance level 0.005 or lower. **: p value < 0.005, ***: p value ≤ 0.001

Feature subset	Features	ρ
Sociodemographic	age	-0.15***
	income	-0.13***
	tabacco days	0.10**
Semantic Interactive	Social: average app duration	0.11**
Semantic Parent	Socials: average app duration	0.10**
Semantic Subcategories	message-boards and forums: average app duration	0.12***
	message-boards and forums: count of apps	0.12***
	parked: count of URLs	0.10**
	parked: fraction of on-events with category	0.10**
	parked: relative duration	0.10**
	parked: total duration	0.11**
Semantic Temporal Subcategories	('A', None) parked: total duration	0.10**
	(None, 'Weekday') parked: total duration	0.11***

Table 14: Desktop - Suicide Risk: PHQ-9-Q9 scores correlation results for the desktop device for significant features at significance level 0.005 or lower. **: p value < 0.005, ***: p value ≤ 0.001

Feature subset	Features	ρ
Sociodemographic	age	-0.15***
	education years	-0.11**
	income	-0.12***
Temporal	('E', 'Weekend'): fraction of URLs	0.11**
	('E', 'Weekend'): relative duration	0.12***

3.5 Classification Analysis

The classification analysis seeks to investigate the potential of the generated internet usage features in depression and suicide risk classification. It pursues this goal by addressing the seven research questions presented in section 1 as part of the second objective of this study.

The classification analysis is performed cross-sectionally with the data from the PHQ-9 interval preceding the baseline survey and the survey responses from the baseline survey. To answer the research questions, the classification analysis is done via an exploratory approach considering all created features and a limited approach where features are pre-selected according to associations found in previous studies. The exploratory approach uses features from all features sets, which are further selected through recursive feature elimination, and explores the classification with four different classifiers (Logistic Regression, Random Forest, SVM-Lin, XGBClassifier). This method is presented in detail in section 3.5.1. The limited approach uses smaller sets of pre-selected features for each feature subset and is limited to two classifiers, Random Forest and SVM-RBF, to replicate results from existing similar studies. This approach is explained more in detail in section 3.5.2. Each approach performs the analysis for three binary splits as the dependent variable.

The splits are two depression score splits, the *Extremes* PHQ-9 split and the *Minimal - Mild Up* PHQ-9 split, and one suicide risk score split, the *No Risk - Suicide Risk* PHQ-9 question 9 (PHQ-9-Q9) split.

The *Extremes* split explores the performance in classifying people with no depression symptoms (PHQ-9 = 0) from people with moderately severe or higher depression severity (PHQ-9 $\geq$ 15), which may have applications in identifying cases that may require more immediate attention or targeted interventions and also shed light on the features that are the most useful in recognizing these two extreme groups.

The *Minimal - Mild Up* split, on the other hand, tests the potential of the created features in recognizing people with no to minimal depression severity (PHQ-9 < 5) from people with mild or greater depression severity (PHQ-9 $\geq$ 5), and has implications on the use of these features in depression prevention.

Similarly, the *No Risk - Suicide risk* split aims to explore the potential of these features in classifying people with no suicide risk symptoms (PHQ-9-Q9 = 0) from people with suicide risk symptoms (PHQ-9-Q9 > 0).

The splits and the counts for each binary outcome per device type is shown in Table 15.

The *Minimal - Mild Up* split and the *No Risk - Suicide Risk* include all selected panelists in the baseline, whereas the *Extremes* split only includes panelists with no depression or moderately severe or higher depression score at baseline, resulting in a significantly smaller sample. As can be observed, the classes for *Minimal - Mild Up* and *Extremes* split are quite balanced for each device, but the *No Risk- Suicide Risk* split is heavily unbalanced in favour of no risk people, as to be expected in the general population.

The feature sets are additionally labelled as online , offline , or online + offline to observe changes in the classification performance for data from different sources. The

Table 15: Splits and number of panelists for each binary outcome. The counts refer to the panelists selected from the baseline survey who have provided desktop (Counts for desktop) or mobile (Counts for mobile) data.

Split	Values	Counts for desktop	Counts for mobile
Minimal - Mild Up	0: PHQ-9 <5	422	390
	1: PHQ-9 ≥ 5	478	483
	Total observations	= 896	= 873
Extremes	0: PHQ-9 = 0	125	117
	1: PHQ-9 ≥ 15	80	91
	Total observations	= 205	= 208
No Risk - Suicide Risk	0: PHQ-9-Q9 = 0	716	679
	1: PHQ-9-Q9 >0	180	194
	Total observations	= 896	= 873

online features sets include all features sets created from IU data (Table 10), which are underlined from this point forward (e.g. Aggregate Volume). The offline feature sets include the sociodemographic features and the demographic features collected in the baseline survey and previously presented in section 3.1.4. These are written in italics from this point forward (e.g. *Sociodemographics*). The online + offline feature sets are the composite feature sets created with IU features and demographic and sociodemographic features. These features sets are wave underlined from this point forward (e.g. Sociodemographics + All IU). The purpose of these groupings is to compare the performance results for IU data (online), against the performance achieved from demographic and sociodemographic features (offline), and observe whether the addition of IU data to demographic and sociodemographic knowledge (online + offline) improves the classification.

3.5.1 Exploratory classification approach

To answer the research questions Q1–Q6 of the second objective, the binary classification methodology presented in Figure 11 is adopted for four different machine learning classifiers and three splits.

The four chosen classifiers included in this exploratory analysis are: Random Forest, XGBClassifier, SVM with linear kernel, and Logistic Regression. These classifiers have different strengths and weaknesses, therefore the analysis is performed for each to gain a more comprehensive understanding of the potential of the created features. Additionally, these classifiers have been selected among others because they allow to use their feature importances or coefficients for the recursive feature elimination function from the `sklearn` python library. The hyper-parameters fixed and tuned for each classifiers are shown in Table 16.

The classification is done independently for each device type and split for the IU feature sets presented in Table 10, with the exception of the Semantic Temporal for the sub-categories because the set is too large and computationally demanding. In addition, classification is performed for the composite set of all internet usage features

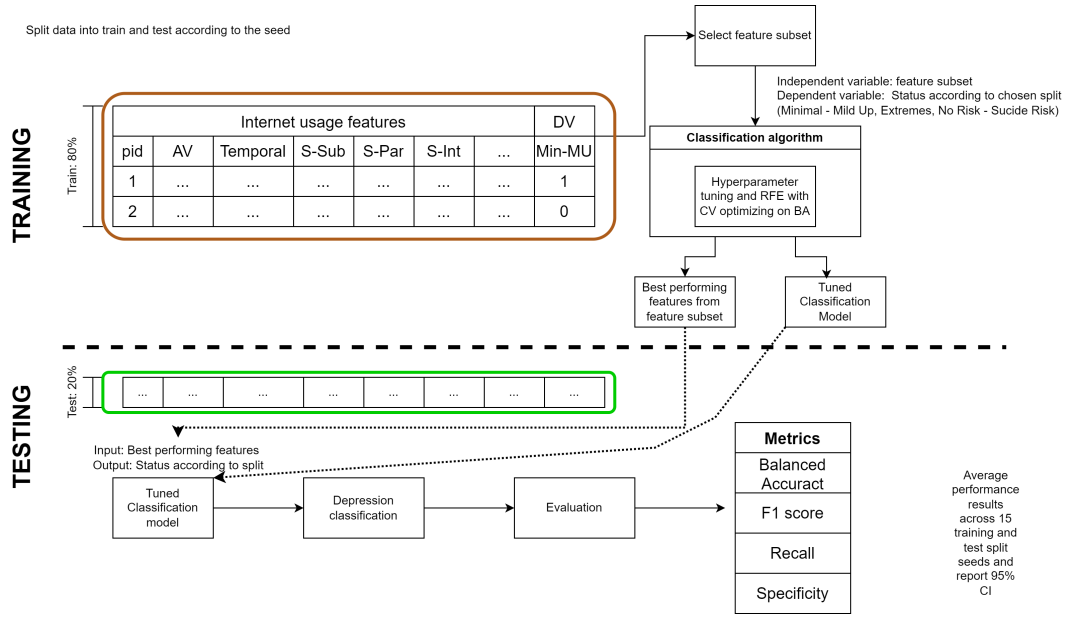


Figure 11: Classification process: For each feature subset, RFECV (recursive feature elimination with cross validation) with hyper-parameter tuning is performed for the model. The selected features and hyper-parameters are used to train the model on the train set (0.8) and tested on the test set (0.2). Results are averaged across 15 train-test stratified splits.

(All IU), the demographic features (age, gender), the sociodemographics features (age, gender, urbanization, education years, income, number of days with tobacco use, political view), the composite set of demographic features and all internet usage features (Demographics + All IU), and lastly the composite set of sociodemographic features and all internet usage features (Sociodemographics + All IU). These are summarized in Table 17.

First, the data is split into a training set (0.8) and a test set (0.2) using one of fifteen shuffling seeds and stratifying on the dependent variable. The dependent variable is a binary variable for the depression status or suicide risk status according to the chosen split (Table 15). The features from the chosen feature set (Table 17) are selected from the train and test set. Then, the model parameters are tuned and the features screened using recursive feature elimination with stratified 5-fold cross validation. The model hyper-parameter tuning and recursive feature elimination with cross validation occur concurrently using Optuna [40], which is a python library for hyper-parameter optimization. Optuna allows to efficiently search large spaces and prune unpromising trials for faster results. A trial in Optuna represents a single execution of the machine learning model with a specific set of hyper-parameters sampled from the hyper-parameter space by a sampler. The TPESampler (Tree-structured Parzen Estimator) is chosen in this study to sample the hyper-parameter space [41] after comparing its performance with the performance of a simple RandomSampler. The TPESampler balances between exploration and exploitation of the hyper-parameter space by using a tree-structured model to guide the search, making it a Bayesian optimization algorithm.

Table 16: Hyper-parameters fixed and tuned for the considered models.

Model	Fixed	Tuned
Logistic regression (LR)	random_state: 42 max_iter: 1000	C: [0.01, 50] class_weight: [None, 'balanced']
SVM with Linear kernel (SVML)	random_state: 42 kernel: linear	C: [0.01, 50] class_weight: [None, 'balanced']
Random Forest (RF)	random_state: 42	n_estimator: [10, 200] max_depth: [2, 100] min_samples_split: [0.1, 1.0] min_samples_lead: [0.1,0.5] max_features: ['sqrt', 'log2'] class_weight: [None, 'balanced']
XGBClassifier (XGB)	objective: binary:logistic eval_metric: logloss booster: gbtree random_state: 42	n_estimators: [50, 200] lambda: [1e-8, 1.0] alpha: [1e-8, 1.0] max_depth: [1,50] scale_pos_weight: [1, #_negative / #_positive] eta: [0.01, 1.0] gamma: [1e-8, 0.05] colsample_bytree: [0.1, 1.0] subsample: [0.5, 1.0] min_child_weight: [1,5]

The number of trials used to tune the hyper-parameters is set to be 100, following to the recommended number of Optuna trials needed for the TPESampler to reasonably prune the hyper-parameter space [41].

In each trial, the TPESampler samples the model hyper-parameters (Table 16) from the hyper-parameter space, and the recursive feature elimination with cross validation (RFECV) selects the features of the feature set (Table 17) that return the best mean cross validation balanced accuracy. In the RFECV, the step size is proportional to the number of features in the set, with a step size of 1 for feature sets under 100 features, 2 for sets under 200 features, 3 for feature sets under 400 and 4 otherwise. The step size is incremented for a faster analysis for the composite sets, but it might lead to important features being sometimes discarded. This is a limitation to take into account. The trial returns the mean cross validation balanced accuracy score and the best performing features with their feature importances. After a hundred Optuna trials, the hyper-parameters and the features selected in the best trial — the trial which returns the best mean cross validation score — are used to train the model and tested on the test set. MinMax scaling is applied during RFECV and to the train and test set. MinMax scaling is selected as the chosen scaling method after comparing its performance with no scaling and standard scaling for the SVM-Lin classifier.

The process is repeated for 15 train-test split seeds to have a comprehensive view of the performance in unseen data. The test results are averaged across the 15 seeds and the 95% confidence interval (CI) from the student T-distribution with 14 degrees of freedom is reported. The training and testing is run on Triton [42], the Aalto University high-performance computing cluster. The training is parallelized across seeds and classifiers on 20-CPU machines. The training duration per seed differs by the classifier and the number of observations in the split. For the *Minimal-Mild Up* and *No*

Table 17: Feature sets included in the exploratory classification analysis. The online source identifies features set from the internet usage (IU) traces (underlined feature sets). The offline source identifies features sets from demographic and sociodemographic information about the panelist, collected in the baseline survey (*italic* feature sets). The online + offline sources identifies feature sets with internet usage features IU and offline panelist information (wave underlined feature sets).

Feature subset	Considered features in RFECV	Source
Aggregate Volume	Aggregate Volume features reported in Table 10	online
Temporal	Temporal features reported in Table 10 for all identified periods (M-A-E-N, Weekday, Weekend, and combinations)	online
Semantic Subcategories	Semantic features reported in Table 10 for the subcategory set (Table 3)	online
Semantic Parent	Semantic features reported in Table 10 for the parent category set (Table 4)	online
Semantic Interactive	Semantic features reported in Table 10 for the interactivity category set (Table 5)	online
Semantic Temporal Parent	Semantic Temporal features reported in Table 10 for the parent category set (Table 4)	online
Semantic Temporal Interactive	Semantic Temporal features reported in Table 10 for the interactivity category set (Table 5)	online
Entropies and KL	Entropies and Kullback-Leiber divergences reported in Table 10	online
All Internet Usage (All IU)	Composite set of all internet usage feature subsets: Aggregate Volume, Temporal, Semantic Subcategories, Semantic Parent, Semantic Interactive, Entropies and KL	online
Demographic	age, gender	offline
Sociodemographic	age, gender, urbanization, tobacco days, politics, income, education years	offline
Demographic + All IU	Composite set of all internet usage features and demographic features	online + offline
Sociodemographic + All IU	Composite set of all internet usage features and sociodemographic features	online + offline

Risk-Suicide Risk splits, the training per seed was the longest for the XGBClassifier (12 hours) and smallest for the Logistic Regression classifier (2 hours). For the *Extremes* split, the training per seed was again the longest for the XGBClassifier (6 hours) and shortest for the Logistic Regression classifier (45 min). The feature importances of the selected features are averaged across the 15 seeds, even if the feature does not get selected in all of the seeds.

The same analysis was also replicated by optimizing on the F1 score instead of the balanced accuracy, which is often the chosen metric in similar studies. Although tuning for the F1 score yielded notably improved recall results, outcomes for specificity were unpromising. Consequently, the decision was made to optimize the classification based on balanced accuracy to achieve a more equitable and balanced classification. Balanced accuracy is the arithmetic mean of sensitivity (true positive rate) and specificity (true negative rate), as shown in Eq. 3. It is particularly useful in scenarios where the classes are imbalanced, which is particularly the case for the *No Risk - Suicide Risk* split.

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2} \tag{3}$$

The results for the exploratory classification approach are shown in section 4.1.1.

3.5.2 Limited classification approach

Similar studies analyzed in section 2.3.1 have used a smaller subset of features in their classification and different classifiers. While recursive feature elimination, as employed in the exploratory classification analysis presented in section 3.5.1, is in general a sound feature selection method, when the number of features is very large in proportion to the number of observations (as in the composite sets All IU, Demographic + All IU and Sociodemographics + All IU), there is the very likely risk of the model selecting sub-optimal features and overfitting on the training set despite the cross-validation. As shown and discussed in the results in section 4.1.1, the composite internet use feature subset (All IU) never performs better than its composing individual feature sets, which might be indicative of poor model generalization. It is reasonable to explore the classification tasks by pre-selecting the features for each feature subset based on known associations found in previous studies.

To limit the number of features, only features from the Aggregate Volume feature set, the Temporal feature set and the Semantic Sub-categories feature set are considered for the analysis. The set of features for each feature set is limited by choosing the most generic and representative of the quantity of internet usage features for each set, which are measures of the duration, the count of URLs and the count of apps. The re-defined limited sets are shown in Table 18.

In the Aggregate Volume set for the mobile data, the number of calls and call duration are also included because they were found to be the most important predictors in a previous study [11]. For the Temporal set, only the time of day periods M-A-E-N are considered. For the Semantic set, the sub-categories are selected based on the features summarized in Table 2. These are the sub-categories *email*, *social-networking*, *chat and messaging* for communication and instant messaging use, *games*, *gambling*, *search-engines and portals* for targeted searches and search engines use, *shopping*, *job-related* for job-related targeted searches, *message-boards and forums* to explore existing associations with suicide ideation, *streaming-media* for streaming. In addition, the *health* subcategory is included on the basis of the psychological literature which claims that depressed individuals may also manifest excessive worry about physical health [17], making it reasonable to include the *health* category in this analysis. Sub-categories instead of the parent categories are used for the limited analysis because they include behaviors, such as instant messaging use, which are lost in the aggregation into parent categorization. The demographic set and the sociodemographic set are the same as in the exploratory classification analysis, while the All Internet Usage (All IU) is the composite of the re-defined limited aggregate volume, temporal and semantic sets.

The models included in the limited analysis are the Random Forest and the SVM with RBF kernel. The RF returned the best performance among 8 other classifiers in a

Table 18: Feature sets included in the limited classification analysis. In addition, the composite set All IU (All IU = {Aggregate Volume, Temporal, Semantic}), Demographic + All IU and Sociodemographic + All IU are created from their composing sets.

Feature set	Included features in limited approach	Device
Demographic	gender	D & M
	age	D & M
Sociodemographic	gender	D & M
	age	D & M
	urbanization	D & M
	education years	D & M
	tobacco days	D & M
	income	D & M
	politics	D & M
Aggregate Volume	Ratio of active days	D & M
	Average daily duration	D & M
	Average daily count of URLs	D & M
	Average daily count of apps	M
Temporal For period *j* in periods: M-A-E-N	Total duration	D & M
	Count of URLs	D & M
	Count of apps	M
Semantic For category *i* in sub-categories: - *shopping* - *chat and instant messaging* - *social-networking* - *health* - *search-engines and portals* - *games* - *gambling* - *streaming-media* - *message-boards and forums* - *job-related*	Total duration	D & M
	Count of URLs	D & M
	Count of apps	M

similar study [11], achieving a AUC score of 0.75 for the classification with mobile usage measures and 0.78 with the classification with mobile usage measures and demographics. The SVM with the RBF kernel was used in another similar study [10] with mobile data using similar features sets, achieving a F1 scores between 0.5-0.8 depending on the feature set.

As per the exploratory classification analysis, hyper-parameter tuning is done using 100 optuna trials with the TPESampler as per the exploratory analysis. Cross

validation (without recursive feature elimination) is done in each optuna trial using 10 stratified folds, 5 more folds than those used in the exploratory analysis thanks to the smaller feature sets, which make computation faster. The tuned hyper-pameters for the RF model and the SVM-RBF model are the same as shown in Table 16, with the exception that the SVM has an RBF kernel. As per the exploratory classification analysis, the performance scores are averaged across 15 train-test split seeds, The results for the limited classification approach are shown in section 4.1.2.

3.6 Longitudinal Analysis: Hierarchical Mixed Effect Models

The correlation analysis (section 3.4) shows statistically significant correlations between several internet but fails to take into consideration potential confounding variables. In contrast, the hierarchical mixed-effects models (HMM) provide a more robust approach by accounting for individual variations, capturing repeated measurements over time, and addressing potential confounding factors. This modeling strategy allows for a nuanced examination of the association between internet usage features and depression or suicide risk outcomes, offering a more comprehensive understanding that goes beyond mere correlations. Hierarchical mixed effects models are statistical models that incorporate both fixed effects and random effects. These models are used to analyze data that exhibit nested or hierarchical structures, where observations are grouped into different levels or clusters. Mixed-effects models are particularly useful when there is a need to account for variability at multiple levels. In this longitudinal study, there is variability for each individual panelist, therefore the panelist ID can be considered as a random effect in the hierarchical model, which translates in practice to linear models with a random intercept for each panelist.

Fixed effects are used to model systematic and non-random influences on the dependent variable. They represent factors for which the goal is to estimate specific, population-level, constant effects in regards to the dependent variable. In the context of this study, the features of interest are the generated internet usage features and the aim is to find the associations of these features with the PHQ-9 and PHQ-9-Q9 scores while also accounting for the fixed effect of sociodemographic features (sociodemographic fixed effects) and individual level variability (panelist random effect). To explore the effect of the internet usage features, the features and survey results from the first three waves of the WebWell study are used. The number of panelists selected in each wave, according to the criteria reported in section 3.1.3, is shown in Table 19. The number of panelists selected in each wave decreases because the selection criteria are not met for that wave or because the panelist dropped from the study. Therefore not all panelists selected at baseline have observations for the second and third wave included in the longitudinal analysis.

Table 19: Number of selected panelist per device type in the first three waves of the WebWell study, according to the panelist selection criteria presented in section 3.1.3.

Wave	Desktop	Mobile
1	896	873
2	798	767
3	589	701

3.6.1 Definition of Hierarchical Mixed Effect Models

The aim of the HMM analysis is not only to identify statistically significant internet usage features but also to assess whether incorporating these features in the analysis

enhances the model fit compared to using a simple baseline model with sociodemographic information. This evaluation provides insights into whether the additional internet usage features contribute significantly to explain the variance in the data, thus informing the model's overall effectiveness in capturing the complexity of the studied phenomena.

First a simple model with sociodemographics fixed effect is run (Model 1). Model complexity is increased by adding features from the aggregate volume feature subset (Model 2), the temporal feature set (Model 3) and the semantic feature subset (Model 4). The general equations for the models tested are defined as follows, where y refers to the feature from the feature subset of interest, y|panelist refers to the random intercept effect on the panelist identifier, $f(\text{feature set}) = c_0 + c_1x_1 + ... + c_nx_n$, where feature set = $\{x_1, ..., x_n\}$ are the surviving variables after the Variance Inflation Factor (VIF) screening for the feature set as is explained in details in section 3.6.2. Additionally, the wave number is included as a fixed effect in all models to take into account potential seasonal effects on depression.

Model 1 := y|pid = f(sociodemographics)

Model 2 := y|pid = f(sociodemographics + aggregate volume)

Model 3 := y|pid = f(sociodemographics + aggregate volume + temporal)

Model 4 := y|pid = f(sociodemographics + aggregate volume + temporal + semantic)

The models are run using the `lme4` R library [43], which uses the Nakagawa et al. [44] method to compute the fixed effect coefficients. A p-value of 0.05 is used to identify statistically significant coefficients among the features for each model. The β coefficients translate to the effect of a unit increase of the feature on the dependent variable, making it challenging to compare the relative importance of the features in the model. For this reason, standardize coefficients are calculated using the Gelman method [45]. The standardized coefficients (β_{std}) are directly comparable and can give more insights about the strength of the features on the dependent variable.

The ICC (Intraclass Correlation Coefficient) is used as a measure of the proportion of total variance in the dependent variable that is attributable to the variation between different panelists. It ranges from 0 to 1, with 0 indicating that there is no variation between panelists, and 1 indicating that all of the variation is between panelists and none within panelists. The marginal R^2 is used as a measure of the proportion of variance explained by the fixed effects included in the hierarchical model. The marginal R^2 ranges from 0 to 1, with 0 indicating that the fixed effects do not explain any variance and 1 indicating that all variance is explained by the fixed effects. the ICC and marginal R^2 are useful in hierarchical fixed effect models to understand the distribution of variance across different levels of the hierarchy and to assess the contribution of fixed effects to explaining the overall variance in the dependent variable.

The more complex models (Model 2 to 4) are compared to the baseline model (Model 1) using the χ^2 one-sided statistical test from the R `anova` function. The χ^2 test assesses whether we can reject the null hypothesis that the simpler model is sufficient and the additional parameters in the more complex model do not significantly improve the fit. If the χ^2 test result from the model comparison is statistically significant, then it can be concluded that the more complex model provides a significantly better fit

compared to the simpler model. If more than one model result to be better than the baseline model from the one-sided test, the χ^2 on the competing models until the best fitting model is found.

In addition to the χ^2 test results, other conventional model selection methods are also taken into consideration in the model comparison. These are Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the log-likelihood value. The AIC and BIC are statistical measures that quantifies the trade-off between the goodness of fit of a statistical model and the complexity of the model, with lower values indicating the best performing model. The main difference between the two is that the BIC incorporates a stronger penalty for the number of parameters than AIC and as a consequence it is likely to penalize the more complex models. On the other hand, the log-likelihood provides a numerical measure of how well a model explains the observed data. A higher log-likelihood indicates a better fit, suggesting that the model is more likely to have generated the observed data. Model results are shown in Appendix G and presented in detail in section 4.2.1 and 4.2.2.

3.6.2 Feature selection for the hierarchical models

For each HMM definition (Models 1 to 4) presented in the previous section, the features need to be selected to limit the number of features to avoid overfitting and avoid multicollinearity between the features to ensure proper model performance. Feature selection is done in two steps. Firstly, only the Aggregate Volume set, the Temporal feature set and the Semantic Sub-categories feature set are considered in the analysis. The set of features for each feature set is limited by choosing the most generic and representative of the quantity of internet usage features for each set exactly as was done for the limited classification analysis shown in section 3.5.2, Table 18.

For each model, the features from the feature sets presented in Table 18 are further screened by doing a variance inflation factor (VIF) analysis to prevent multicollinarity among the features. Multicollinearity occurs when predictor variables in a regression model are correlated, leading to challenges in estimating the individual effects of the predictors. The variance inflation factor is a measure that quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity. For a predictor i in the predictors set of length n, the VIF value for feature i VIF_i is shown in Eq. 4, where R_i^2 is the R^2 value obtained by regressing the i-th predictor against all the other predictors in the model.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{4}$$

A VIF value of 1 indicates no multicollinearity. The choice of the VIF threshold is sometimes arbitrary and context dependent, but generally a variance inflation factor of 5 has been used in the literature to show that there are acceptable degrees of multicollinearity among the features. In this case though, a more stringent value is used because different quantity measurements (duration, URL count and app count) for the same subject (e.g. gaming activity) are likely to be correlated, and ideally only the ones that explain the most variance should be included, otherwise there is

the risk that the effect of the two correlated features cancel each other in the analysis. Screening the features using a more stringent VIF value ensures that there is negligible multicollinearity among the features, which improves the stability in the estimates of the regression coefficients and allows for better interpretability of individual predictor effects. A VIF value ≥ 2.5 has been shown in the literature to indicate considerable collinearity [46]. Therefore, a smaller VIF value of 1.5 is used in this study to iteratively eliminate features to use in the model until all selected features have VIF value under 1.5. The model is then run with the surviving features. The final sets of features included for each model after the VIF screening is shown in the result tables presented in Appendix G. As can be observed, several of the features presented in the less complex models are dropped because of multicollinearity with other features considered in the more complex models.

4 Results

In this section, the outcomes of the classification analyses outlined in sections 3.5.1–3.5.2 are presented and discussed in section 4.1. These analyses focus on the second objective of the study, aiming to explore how internet usage features contribute to depression and suicide risk classification. Additionally, the findings from the HMM analysis introduced in section 3.6 are reported and discussed in section 4.2, addressing the third objective of the study by investigating the intricate associations between internet usage patterns and depression and suicide risk.

4.1 Classification analysis

The following section presents the results from the classification analyses. Section 4.1.1 reports a summary of the results from the exploratory classification analysis, reported fully in Appendix E. Section 4.1.2 reports a summary of the results from the limited classification analysis, reported fully in Appendix F. The most important features by mean importance value are illustrated in section 4.1.3. Lastly, section 4.1.4 summarizes and compares the results from the classification analyses in view of the research questions first presented in section 3.5, and contrasts the findings with the existing literature.

4.1.1 Exploratory classification

The exploratory classification analysis relied on recursive feature elimination to select the best performing features in each feature set. The best results per feature set and device type are summarized in Table 20, where the best balanced accuracy from the best performing model is reported for each feature set. The full results are shown in Appendix E, in Tables E1–E2 for the *Extremes* split, Tables E3–E4 for the *Minimal - Mild Up* split, and Tables E5–E6 for the *No Risk - Suicide Risk* split, for mobile and desktop devices respectively.

The results of the exploratory classification analysis clearly show that the best performing feature set is the *Sociodemographic* feature set for all splits, indicating that sociodemographic knowledge has higher potential in depression classification than the explored internet usage feature sets. This is consistent with the fact that depression is highly correlated with sociodemographic factors such as income, age, gender, and tobacco use, as shown in the correlation analysis in section 3.4 and in the psychological literature [17][13][20].

Moreover, the results show that the addition of the IU sets to the sociodemographic or demographic sets (online + offline : Demographic + All IU and the Sociodemographic + All IU sets), always outperforms the accuracy achieved from IU sets (online), but does not outperform the *Demographic* or *Sociodemographic* sets (offline) across the splits and devices. This could be indicative of poor feature selection in the RFECV procedure and consequent poor model generalization due to the increase in the feature set size, or that the addition of internet usage features has actually a confounding effect on the classification. The composite set of all internet

usage features (All IU) never outperforms its composing IU features sets (Aggregate Volume, ..., Temporal Semantic), which implies that more privacy intrusiveness does not always relate to better performance in this analysis, perhaps as a consequence of the feature selection method. From the results, there is no apparent classifiers that outperforms the others for all features sets.

Table 20: Best performance scores from the exploratory classification analysis using RFECV for feature selection. For each split, device and feature set, the mean balanced accuracy and 95% CI (BA) from the best performing classifier (C) is reported. See Table 10 for the features considered in the RFECV for each feature set. Results are the average of 15 train-test split seeds, and the 95% CI are calculated from the t-statistic with df=14. Cells colored in blue point to the best performing feature set for the online only feature sets (underlined). Cells colored in orange point to the best performing feature set for the offline only feature sets (*italic*). Cells colored in green point to the best performing feature set for the online + offline feature sets (wave underlined). Splits: *Extremes* (PHQ-9 = 0 - PHQ-9 ≥ 15), *Minimal - Mild Up* (PHQ-9 < 5 - PHQ-9 ≥ 5), *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0).

Split	**Extremes**				**Minimal - Mild Up**				**No Risk - Suicide Risk**			
Device	**Desktop**		**Mobile**		**Desktop**		**Mobile**		**Desktop**		**Mobile**	
Best score/Classifier	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**
Aggregate Volume	0.57 ± 0.05	RF	0.53 ± 0.04	LR/RF	0.56 ± 0.02	LR	0.55 ± 0.02	LR	0.50 ± 0.02	XGB	0.52 ± 0.03	SVML
Temporal	0.54 ± 0.04	LR	0.55 ± 0.05	SVML	0.55 ± 0.01	RF	0.52 ± 0.02	RF/SVML/XGB	0.54 ± 0.02	LR	0.51 ± 0.02	RF
Semantic Subcategories	0.55 ± 0.04	XGB	0.56 ± 0.03	RF	0.53 ± 0.02	LR/RF	0.53 ± 0.02	LR	0.52 ± 0.03	RF	0.50 ± 0.02	RF/SVML
Semantic Parent	0.54 ± 0.03	LR	0.53 ± 0.03	RF/SVML	0.53 ± 0.01	LR/SVML	0.54 ± 0.02	LR	0.51 ± 0.03	LR/RF	0.51 ± 0.02	LR
Semantic Interactive	0.53 ± 0.04	XGB	0.59 ± 0.04	LR	0.54 ± 0.02	RF	0.55 ± 0.02	RF	0.50 ± 0.02	XGB	0.51 ± 0.03	RF
Entropies and KL	0.50 ± 0.05	SVML	0.51 ± 0.05	XGB	0.55 ± 0.02	SVML	0.50 ± 0.01	RF	0.50 ± 0.02	LR	0.52 ± 0.03	XGB
Temporal Semantic Interactive	0.57 ± 0.05	XGB	0.58 ± 0.03	RF	0.55 ± 0.02	RF	0.54 ± 0.01	SVML	0.52 ± 0.02	SVML	0.52 ± 0.02	RF
Temporal Semantic Parent	0.60 ± 0.03	XGB	0.57 ± 0.03	XGB	0.55 ± 0.02	RF	0.52 ± 0.01	RF	0.51 ± 0.02	XGB	0.51 ± 0.02	LR/SVML
All IU	0.58 ± 0.04	XGB	0.56 ± 0.04	RF	0.55 ± 0.02	RF	0.53 ± 0.03	XGB	0.52 ± 0.02	RF	0.51 ± 0.02	LR
Demographic	0.64 ± 0.05	LR	0.64 ± 0.05	XGB	0.59 ± 0.02	XGB	0.61 ± 0.02	LR	0.59 ± 0.02	RF	0.58 ± 0.03	RF
Sociodemographic	0.72 ± 0.04	RF	0.73 ± 0.03	SVML	0.60 ± 0.02	LR/SVML/XGB	0.63 ± 0.02	LR/SVML	0.63 ± 0.02	XGB	0.58 ± 0.03	SVML/XGB
Demographic + All IU	0.64 ± 0.04	XGB	0.58 ± 0.04	LR	0.57 ± 0.01	LR	0.57 ± 0.02	LR/SVML	0.55 ± 0.03	SVML	0.53 ± 0.02	LR
Sociodemographic + All IU	0.64 ± 0.03	XGB	0.64 ± 0.04	LR	0.60 ± 0.02	LR	0.60 ± 0.02	LR	0.56 ± 0.02	LR/SVML	0.54 ± 0.02	LR

4.1.2 Limited classification

The limited classification analysis relied on results from existing literature to pre-select features to use in each set for the classification. The best results per feature set and device type are summarized in Table 21, where the best balanced accuracy from the best performing model is reported for each feature set. The full results are shown in Appendix F, in Tables F1–F2 for the *Extreme* split, Tables F3–F4 for the *Minimal - Mild Up* split, and Tables F5–F6 for the *No Risk - Suicide Risk* split, for mobile and desktop devices respectively.

The results of the limited classification analysis, similarly to the the results from the exploratory classification analysis, show that the best performing feature set is the *Sociodemographic* (offline) feature set for all splits, and that the addition of the IU sets to the sociodemographic or demographic sets (online + offline : Demographic + All IU and theSociodemographic + All IU sets) does not outperform the *Demographic* or *Sociodemographic* sets (offline). Additionally, as per the exploratory classification analysis, the composite set of all internet usage features (All IU) never outperforms its composing IU features sets (Aggregate Volume, ..., Temporal Semantic), which implies that more privacy intrusiveness does not always translate to better performance in this analysis. While it was speculated that the under-performance of the composite sets (Sociodemographics + All IU, Demographics + All IU, All IU) versus their composing sets could be indicative of poor feature selection in the RFECV procedure and poor model generalization in the exploratory analysis due to increased model complexity, this hypothesis is less valid in the limited analysis because the features in each feature sets are pre-selected and the composite features sets (Sociodemographics + All IU, Demographics + All IU, All IU) are much smaller. Additionally, as per the exploratory analysis, regularization techniques are applied to the classifiers which should prevent them from generalizing poorly, something which is also tackled by the higher number of folds (10 versus the 5 used in the exploratory analysis) in the cross-validation. Nevertheless, it is still possible that that the added model complexity due to the increased features set size is preventing the models from generalizing. Future analysis should integrate RFECV in the composite feature sets, to observe whether additional feature selection helps the models detaining only the most important features in the composite sets and improve model performance.

Table 21: Best performance scores from the limited classification analysis using pre-selected features on the basis of existing literature. For each split, device and feature set, the mean balanced accuracy and 95% CI (BA) from the best performing classifier (C) is reported. See Table 18 for the features included in each features set. Results are the average of 15 train-test split seeds, and the 95% CI are calculated from the t-statistic with df=14. Cells colored in blue point to the best performing feature set for the online only feature sets (underlined). Cells colored in orange point to the best performing feature set for the offline only feature sets (*italic*). Cells colored in green point to the best performing feature set for the online + offline feature sets (wave underlined). Splits: *Extremes* (PHQ-9 = 0 - PHQ-9 ≥ 15), *Minimal - Mild Up* (PHQ-9 < 5 - PHQ-9 ≥ 5), *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0).

Split	**Extremes**				**Minimal - Mild up**				**No Risk - Suicide Risk**			
Device	**Desktop**		**Mobile**		**Desktop**		**Mobile**		**Desktop**		**Mobile**	
Best score/Classifier	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**	**BA**	**C**
Aggregate Volume	0.61 ± 0.04	SVM-RBF	0.57 ± 0.02	SVM-RBF	0.56 ± 0.02	SVM-RBF	0.54 ± 0.02	RF	0.50 ± 0.02	SVM-RBF	0.49 ± 0.01	SVM-RBF
Temporal	0.52 ± 0.03	SVM-RBF	0.52 ± 0.02	RF	0.55 ± 0.02	RF	0.52 ± 0.01	SVM-RBF	0.49 ± 0.02	SVM-RBF	0.49 ± 0.02	RF
Semantic	0.51 ± 0.03	RF	0.57 ± 0.03	RF	0.53 ± 0.01	SVM-RBF	0.54 ± 0.01	RF	0.49 ± 0.01	SVM-RBF	0.49 ± 0.02	SVM-RBF
All IU	0.58 ± 0.04	SVM-RBF	0.54 ± 0.03	RF	0.55 ± 0.01	SVM-RBF/RF	0.54 ± 0.01	SVM-RBF	0.50 ± 0.01	SVM-RBF	0.50 ± 0.01	SVM-RBF
Demographic	0.65 ± 0.04	SVM-RBF	0.64 ± 0.03	RF	0.59 ± 0.01	SVM-RBF	0.60 ± 0.02	RF	0.59 ± 0.02	RF	0.58 ± 0.03	RF
Sociodemographic	0.71 ± 0.05	RF	0.70 ± 0.03	SVM-RBF	0.60 ± 0.02	RF	0.62 ± 0.02	RF	0.61 ± 0.02	RF	0.59 ± 0.02	SVM-RBF
Demographic + All IU	0.64 ± 0.04	SVM-RBF	0.60 ± 0.03	SVM-RBF	0.57 ± 0.02	RF	0.60 ± 0.01	SVM-RBF	0.55 ± 0.02	RF	0.52 ± 0.01	SVM-RBF
Sociodemographic + All IU	0.66 ± 0.05	RF	0.65 ± 0.03	SVM-RBF	0.59 ± 0.02	SVM-RBF	0.61 ± 0.01	SVM-RBF	0.57 ± 0.03	SVM-RBF	0.53 ± 0.02	RF

4.1.3 Feature Importances

As described in section 3.5, the feature importances or model coefficients of the selected features are collected for each seed during the training process. The importances refer to the trained model coefficients (for XGB, LR, SVM-L) or feature importances (RF) of the selected features. For clarity, they are referred as 'importances' in this section. This section explores the features with the greatest importance in the classification analysis to address the second objective of this work, that is to identify the best performing features in depression or suicide risk classification.

As shown in Tables 20–21 presented in the previous sections, the Sociodemographic + All IU is the best performing feature set with IU features. It is also the most complete feature set, even if it never outperforms the balanced accuracy achieved by its composing *Sociodemographic* set. Nevertheless, it is explored in this section because the aim is to understand which internet usage features, in addition to or besides the sociodemographic ones, have been the most useful in the classification. The top 20 features are the features whose importance is the highest after averaging across the number of test-train split seeds. For the limited analysis, the importances are accessible only for the Random Forest classifier, despite the SVM-RBF often outperforming the RF model in the Sociodemographic + All IU set (see Table 21). The importances for Sociodemographic + All IU set in the desktop *Extremes* split for the limited classification analysis are shown in Figure 12. The importances for the best performing models with the Sociodemographic + All IU set are shown in Figure 13a–13b for the exploratory classification analysis.

The importances are not comparable across models, but they given an indication of which feature have been selected the most across seeds and their relative importance for the best performing model. What can be observed across splits for each classification analysis (exploratory and limited) is that sociodemographic features, most importantly income and age, appear repeatedly in the most important features, often ranking first or second in average importance. This reinforces the findings from the correlation analysis in section 3.4, which show that sociodemographic features demonstrate the highest correlation coefficients with depression and suicide risk classification. Regarding the IU features, it is challenging to make general takeaways about the most important features across splits and devices. For the *Extremes* split, the most important IU feature is the total duration spent on *email* for desktop devices (Limited analysis, Figure 12), and the average app duration spent on

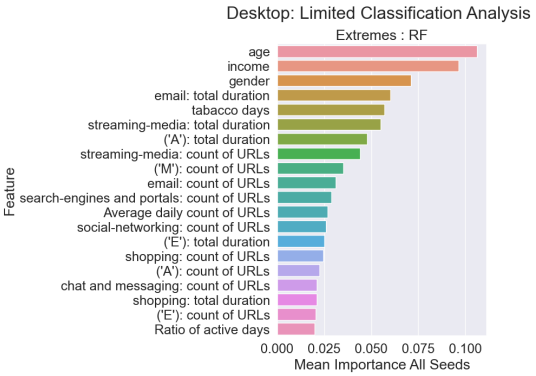


Figure 12: Top features by mean features importance for the limited classification results on the Sociodemographic + All IU set on the desktop devices for the Extreme split. Mean balanced accuracy: 0.66. Model: Random Forest

(a) Desktop: Mean feature importance of top features for the best classifiers in each split (BA = 0.64, 0.60, 0.56).

(b) Mobile: Mean feature importance of top features and best classifiers in each split (BA = 0.64, 0.60, 0.54).

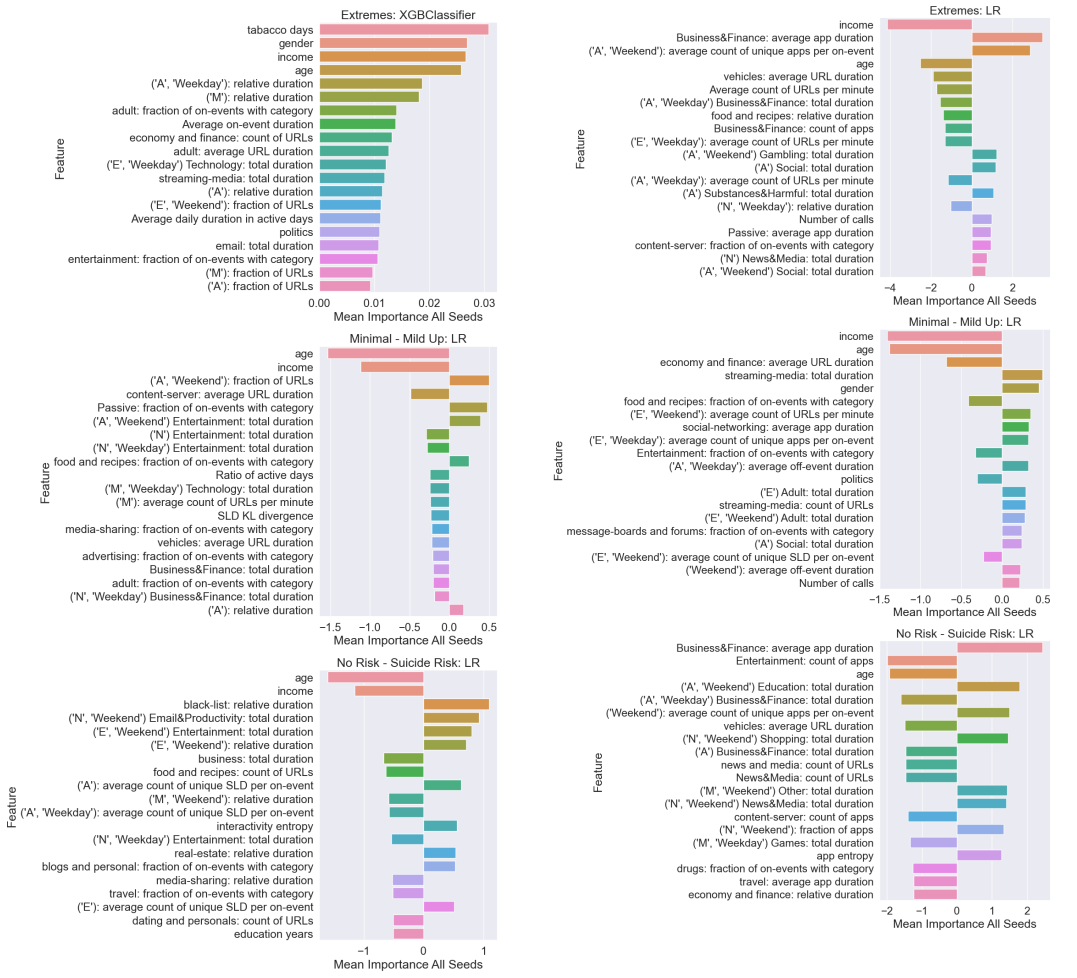
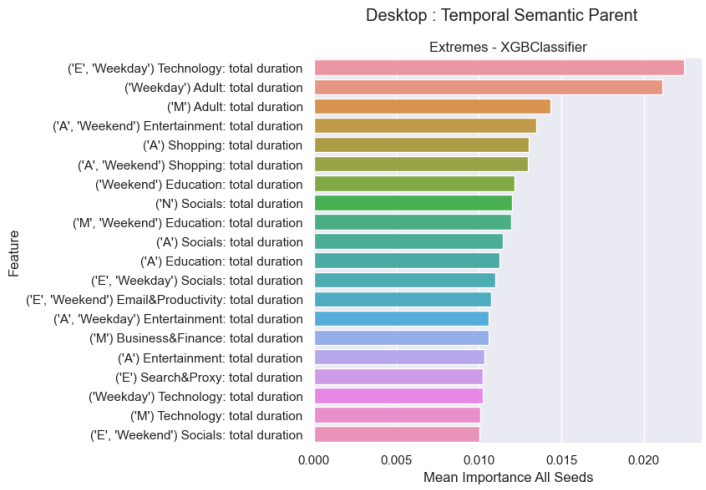


Figure 13: Top features by mean features importance for the exploratory classification results on the Sociodemographic + All IU set (left: desktop, right: mobile) for each split (top: *Extremes* (PHQ-9 = 0 - PHQ-9 ≥ 15), middle: *Minimal - Mild Up* (PHQ-9 < 5 - PHQ-9 ≥ 5), bottom: *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0). BA refers to the mean balanced accuracy (order: *Extremes*, *Minimal - Mild Up*, *No Risk - Suicide Risk*). Only the results from the best performing classifiers are shown.

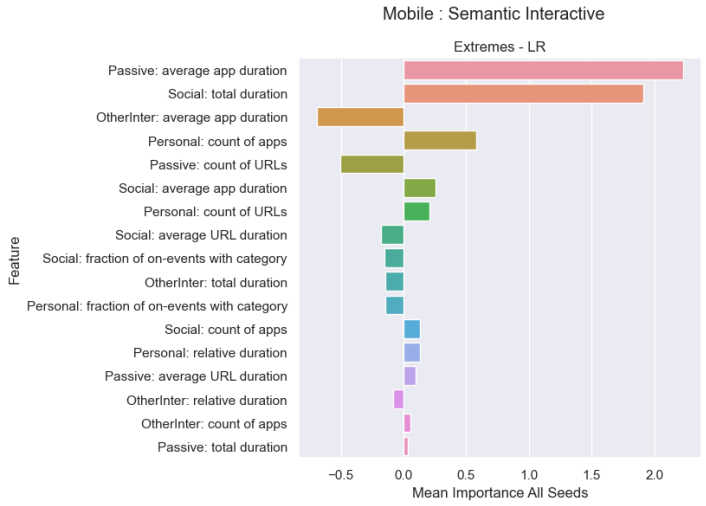
Business&Finance apps for mobile devices (Exploratory analysis, Figure 13b, top). For the *Minimal - Mild up* split, the most important IU feature for desktop devices is the fraction of URLs on weekend afternoons (Exploratory analysis, Figure 13a, middle), while for mobile devices the most important IU features is the average URL

duration of *economy and finance* URLs (Exploratory analysis, Figure 13b, middle). Lastly, for the *No Risk - Suicide Risk* split, the most important IU feature is the relative duration spent on *black-listed* content for desktop devices (Exploratory analysis, Figure 13a, bottom) and the average *Business&Finance* app view duration for mobile devices (Exploratory analysis, Figure 13b, bottom).

In addition to the selected features in the Sociodemographic + All IU sets, it is also worth exploring the top features in the Temporal Semantic Parent set for the *Extremes* split on desktop device and the top features in the Semantic Interactive set in the *Extremes* split on mobile for the exploratory analysis. These sets achieved a balanced accuracy of 0.60 and 0.59 respectively (Table 20), demonstrating the highest potential for classification among the IU only (online) feature sets for the *Extremes* split on the exploratory analysis, making it worth observing which features have been selected. The top mean feature importance for the Temporal Semantic Parent set for desktop and the Temporal Interactive set for mobile are shown in Figure 14a–14b respectively. Figure 14a shows that for the desktop devices, evening weekday viewing of *Technology* content is the most important predictor in the *Extremes* split, followed by *Adult* content viewing, Entertainment, Shopping, and Socials viewing in the night. Figure 14b shows that for mobile devices, the average time spent on apps labelled as passive interactive (e.g *streaming-media, entertainment*, please refer to Table 5 for the definition of the interactivity categories) is the most important positive predictor in the *Extremes* split, followed by the total time spent on social interactive apps (e.g *media-sharing, chat and messaging, social-networking, message-boards and forums, email*), and the number of app views labelled as personal interactive (*games, gambling, shopping, productivity, survey*).



(a) Desktop: mean feature importance of top features for the desktop device in the Temporal Semantic Parent feature set for the *Extremes* split (BA = 0.60).



(b) Mobile: mean feature importance of top features for the mobile device in the Semantic Interactive feature set for the *Extremes* split (BA = 0.59).

Figure 14: Top features by mean features importance for the exploratory classification results on the Temporal Semantic Parent set for desktop devices (top) and the Semantic Interactive set for mobile devices (bottom) for the *Extremes* (PHQ-9 = 0 - PHQ-9 ≥ 15) split. BA refers to the mean balanced accuracy. Only the results from the best performing classifiers are shown.

4.1.4 Summary of classification results

The best results from the classification analyses, both exploratory and limited, are summarized in Table 22. For each of the online, offline, and online + offline feature sets, the best performing result from the best performing set, classifier, feature selection method (exploratory or limited) are reported for each split and device.

Table 22: Best performance on the binary classification analysis: summary of best average balanced accuracy results for each split by device and data source type. The best performing feature set is reported in the feature set field (see Tables 10–18 for the feature sets definitions), in addition to a letter indicating the type of feature selection method (**E**: Exploratory analysis with features selected from RFECV, **L**: Limited analysis with features pre-selected from existing literature) and the best performing model. Splits: *Extremes* (PHQ-9 = 0 - PHQ-9 ≥ 15), *Minimal - Mild Up* (PHQ-9 < 5 - PHQ-9 ≥ 5), *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0). Source: online (considered: Aggregate Volume, Temporal, Semantic Subcategories, ..., All IU), offline (*Demographics*, *Sociodemographics*), online + offline (Demographic + All IU, Sociodemographic + All IU)

Source	Online				Offline				Online + Offline			
Device	Desktop		Mobile		Desktop		Mobile		Desktop		Mobile	
Best score/Feature set	BA	Feature set	BA	Feature set	BA	Feature set	BA	Feature set	BA	Feature set	BA	Feature set
Extremes	0.61 ± 0.04	Aggregate Volume (**L**, *SVM-RBF*)	0.59 ± 0.04	Semantic Parent (**E**, *LR*)	0.72 ± 0.04	*Socio-demographic* (**E**, *RF*)	0.73 ± 0.03	*Socio-demographic* (**E**, *SVML*)	0.66 ± 0.05	Socio-demographic + All IU (**L**, *RF*)	0.65 ± 0.04	Socio-demographic + All IU (**L**, *SVM-RBF*)
Minimal - Mild Up	0.56 ± 0.02	Aggregate Volume (**L**, *SVM-RBF*)	0.55 ± 0.02	Aggregate Volume (**E**, *LR*)	0.60 ± 0.02	*Socio-demographic* (**E/L**, *RF*)	0.63 ± 0.02	*Socio-demographic* (**E**, *LR/SVML*)	0.60 ± 0.02	Socio-demographic + All IU (**E**, *LR*)	0.61 ± 0.01	Socio-demographic + All IU (**L**, *SVM-RBF*)
No Risk - Suicide Risk	0.54 ± 0.02	Temporal (**E**, *LR*)	0.52 ± 0.03	Aggregate Volume (**E**, *LR*)	0.63 ± 0.02	*Socio-demographic* (**E**, *XGB*)	0.59 ± 0.02	*Socio-demographic* (**L**, *SVM-RBF*)	0.57 ± 0.03	Socio-demographic + All IU (**L**, *SVM-RBF*)	0.54 ± 0.02	Socio-demographic + All IU (**E**, *LR*)

The research questions proposed under the second objective of this study in section 1 are answered in view of the results reported in Tables 21–20 and summarized in Table 22.

- **Q1:** *How does the performance differ across device type (desktop and mobile)? Which data from which device is more insightful for depression classification?*

  Performance across devices is very similar, but on average data from desktop devices achieves better average balanced accuracy scores.

- **Q2:** *How does the performance differ for each of the created internet usage feature subsets? Does more privacy intrusiveness relate to better performance?*

  For the online only feature sets, lower intrusiveness feature sets generally outperform higher intrusiveness feature set for the best performing classifier. The Aggregate Volume feature set is the most promising online only feature set on average, indicating that information on the aggregate quantity of internet volume is on average a better predictor than temporal and semantic internet usage information in this analysis.

- **Q3:** *How does the performance differ between using internet usage features only (online features), demographic or sociodemographic features only (offline features), and internet usage plus demographic or sociodemographic features (online + offline features)? Is there an improvement in results obtained by including internet usage features compared to the performance achievable with demographic or sociodemographic information?*

  The offline feature sets always outperform the online only feature sets. The online + offline feature sets outperform the online feature sets, but do not outperform the offline only feature sets, indicating that demographic and sociodemographic information is more important for depression status classification for the observed populations across splits.

- **Q4:** *How does the performance differ between classifying people with mild or lower depression severity (PHQ-9 $<$ 5) from people with mild or greater depression severity (PHQ-9 $\geq$ 5) versus classifying people with no depression symptoms (PHQ-9 = 0) from people with moderately severe or higher depression severity (PHQ-9 $\geq$ 15)? Can this technology be useful in early depression diagnosis?*

  The average performance of the *Extremes* split (PHQ-9 = 0 - PHQ-9 $\geq$ 15) is higher than the performance of the *Minimal - Mild Up* split (PHQ-9 = 0 - PHQ-9 $\geq$ 15) for all feature sets, including the offline only sets. Overall, it is easier to distinguish people with no depression severity from people with high depression severity, than people with minimal or lower depression severity form people with mild or higher depression severity. The best accuracy achieved in on the *Minimal - Mild Up* split using online + offline feature sets, which could be used for early depression diagnosis, is 0.61 (see Table 22).

- **Q5:** *What is the performance for classifying people with no suicide risk (PHQ-9-Q9 = 0) from people with suicide risk (PHQ-9-Q9 > 0)? Can this technology be used in early suicide risk detection?*

  The performance of IU feature sets for suicide risk detection is very poor, indicating that the created internet usage features are not particularly useful for suicide risk diagnosis. The performance improves when adding sociodemographic features, but never outperforms the performance from sociodemographic features alone.

- **Q6:** *What are the selected features for the IU feature sets with the highest importance? Which internet behaviours are the most useful in depression classification?*

  For the Sociodemographics + All IU feature set, the most important features are generally sociodemographic features. Income and age are selected in the exploratory analysis in almost all splits. For the *Extremes* split, the most important IU feature is the total duration spent on *email* for desktop devices (Limited analysis, Figure 12), while for mobile devices the most import IU

feature is the average app duration spent on *Business&Finance* apps (Exploratory analysis, Figure 13b, top). For the *Minimal-Mild up* analysis, the most important IU feature from desktop devices is the fraction of URLs on weekend afternoons (Exploratory analysis, Figure 13a, middle), while for desktop devices the most important IU feature is the average URL duration of *economy and finance* URLs (Exploratory analysis, Figure 13b, middle). Lastly, for the *No Risk - Suicide Risk* split, the most important IU feature is the relative duration spent on *black-listed* content for desktop devices (Exploratory analysis, Figure 13a, bottom) and the average *Business&Finance* app view duration for mobile devices (Exploratory analysis, Figure 13b, bottom).

- **Q7:** *How do the results compare to those achieved in similar studies, when using similar feature sets?*

  C.Yue et al. [10] were able to achieve an F1 of about 0.71 (iOS) and 0.80 (Android) from a feature set including volume, temporal and semantic information, which is closely comparable to the All IU set used in the limited analysis (Table 18). In comparison, the best balanced accuracy achieved in this study for mobile data from IU features is 0.59 for the *Extremes* split and 0.55 for the *Minimal-Mild up* split (Table 22). C.Yue et al. also find that adding more intrusive internet usage features, specifically semantic and temporal features, returns better performance, which is not always the case in this analysis as previously discussed. The main difference between this study and their study is that their population comprised 145 university students with depression status classified by a professional psychologist, making it challenging to compare their analysis with any of the splits (*Extremes*, *Minimal-Mild Up*, *No Risk - Suicide Risk*) and the significantly larger and more heterogeneous population of this study. Additionally, it could also be argued that their classification is less sensitive to the issues with survey self-assessment, which are perhaps problematic in this study relying on self-administered PHQ-9 questionnaires.

  Razavi et al. [11] conducted an analysis on mobile devices that is comparable to the *Minimal - Mild Up* limited classification analysis (Table 21), with the exception that they had detailed access to the number of calls and messages received and sent by each participants and the number of contacts saved in their device, and that their classes were heavily unbalanced (3 to 1) in favour of the minimal class. Their population size is 412 participants (equal split across females and males, average age of 40 with standard deviation of 18.87), significantly larger than the populations observed in other studies but still less than half the size of the one explored in this work. Their methodology is very similar to the one used in the limited classification analysis, and their best performing classifier was the Random Forest, which achieves a BA = 0.76 on mobile usage features only and BA = 0.83 on mobile usage + demographic features. In this study, the best accuracy achieved with the IU features for mobile devices was BA = 0.55 (Table 22 with LR) and the best accuracy achieved with internet usage and demographic features was BA = 0.60 (Table 21 with

SVM-RBF). Their feature importance analysis reveals that the number of calls made daily, the average time spent online, and the number of contacts saved on the device are the most important features in their classification analysis. The first and third features are not directly available from the traces used in this study. The number of calls presented in Tables 10–18 is only a proxy and does not distinguish between the number of calls made or received, and doesn't include calls made from WhatsApp or other messaging apps. Regardless, the number of calls is still among the top 20 features for mobile by average importance (Figure 13b, middle).

To conclude, this study shows results that are less promising than those achieved in similar studies. It can be speculated that this is because this study uses a larger more heterogeneous population representative of the general demographic, and that perhaps it is because it does not include features, such as detailed information about the number of calls made and received and the number of contacts in the device, that have been shown to have the greatest potential in depression classification.

In conclusion, the results from the classification analysis reveal that sociodemographic information, specifically income and age information, detain the highest potential in classifying both depression and suicide risk, and that adding internet usage features from mobile and desktop devices does not improve the performance. Regardless, data from desktop devices appears to be slightly more promising than the explored data from mobile devices, and recognizing people with no depression from people with high depression severity (*Extremes* split) achieves the best accuracy across devices.

The classification on IU data only reveals that more privacy intrusive features do not always translate to better performance, with few exceptions, and that information about the total volume of internet usage (Aggregate Volume set) returns on average the best performance. This is in contrast with a previous study [10], which showed that semantic and temporal information always perform better for their population. The best results with IU data are achieved when adding sociodemographics features (Sociodemographic + All IU set) for both mobile (*Extremes*: 0.65 BA, *Minimal - Mild Up*: 0.61, *No Risk - Suicide Risk*: 0.54) and desktop devices (*Extremes*: 0.66, *Minimal - Mild Up*: 0.60, *No Risk - Suicide Risk*: 0.54), although never outperforming the balanced accuracy achieved from the *Sociodemographic* set alone. In the exploratory analysis, it is possible that this is the results of poor feature selection in the RFECV due to the large size of the composite feature sets and increased model complexity. The validity of this argument diminishes in the context of the limited analysis, characterized by a significantly smaller feature sets. To address this, further investigation, incorporating additional feature selection methods such as Recursive Feature Elimination (RFE) for the limited analysis, along with a detailed examination of training and testing performance, is warranted. This exploration could help determine whether the observed issues are attributable to poor model generalization or the confounding effects of certain included IU features. Given the use of regulation techniques in most classifiers, and the thorough hyper-parameter tuning with cross validation, it is possible that

the reason is due to a possible confounding effect of internet usage features on the classification when observing the general population.

4.2 Hierarchical Mixed Effect Models

This section discusses the results from the hierarchical mixed effect models (HMMs) analysis. Section 4.2.1 presents the results for the analysis on the PHQ-9 depression score as the dependent variable. Section 4.2.2 presents the results for the suicide risk PHQ-9-Q9 score as the dependent variable. The full tables are reported in Appendix G. For each dependent variable and device, the Sociodemographic model (Model 1) is considered the be the baseline model. To observe whether the addition of the fixed effects in the more complex models (Model 2: Sociodemographic + Aggregate Volume, Model 3: Sociodemographic + Aggregate Volume + Temporal, Model 4: Sociodemographic + Aggregate Volume + Temporal + Semantic) is reasonable, a one sided χ^2 test between the Sociodemographic (Model 1) baseline and the more complex models (Model 2-4) is performed. If one of the more complex model is significantly better at capturing the data than the simpler baseline model, then the p-value of the χ^2 statistic from the one-sided test should be significant. In addition to the one-sided χ^2 test, the models are compared by observing the BIC, AIC and log-likelihood values, with smaller BIC, AIC and higher log-likelihood indicating a better fit.

4.2.1 Depression

For each of the specified model presented in section G, the model results with the PHQ-9 depression score as the dependent variable for the mobile and desktop devices are shown in Tables G1–G2 in Appendix G respectively.

For the mobile data (Table G1), the more complete model (Model 4: Sociodemographic + Aggregate Volume + Temporal + Semantic) shows that age and income have statistically significant negative associations with depression, while the number of days with tobacco use, being a female, the total count of app views during the night, the total duration spent on *chat and messaging* platforms, and the total count of *job-related* URL queries have statistically significant positive associations with depression.

It can be observed that every extra app view in the night increases the PHQ-9 score by 0.001 (95% CI (0.000–0.001), $\beta_{\text{std}} = 0.042$, $P < 0.01$), every extra minute spent on *chat and messaging* platforms increases the PHQ-9 score by 0.001 (95% CI (0.000–0.002), $\beta_{\text{std}} = 0.046$, $P < 0.01$), and that every extra *job-related* URL query increases the PHQ-9 score by 0.024 (95% CI (0.011–0.037), $\beta_{\text{std}} = 0.075$, $P < 0.001$). The standardized β_{std}s reveal that the number of the *job-related* URLs has the greatest impact on the PHQ-9 score among the statistically significant internet usage features.

Model 2 (Sociodemographic + Aggregate Volume) additionally shows that the average daily count of apps has a statistically significant positive association with depression, with every extra daily app view increasing the PHQ-9 score by 0.003 (95% CI (0.001–0.004), $\beta_{\text{std}} = 0.066$, $P < 0.01$). The average daily count of apps

is not included in Model 3–4 because it is eliminated by the VIF threshold to avoid multicollinearity.

The ICC across all models is 0.78, indicating that the individual panelist random effect explain 78% of the total variance explained by each model. In the more complex model (Model 4), the marginal R^2 is 0.1, highlighting that the fixed effects explain 10% of the total variance explained by the model. The high ICC and low marginal R^2 show that the PHQ-9 scores are mostly heavily dependent on the individual panelist random effect.

Model comparison results (Table 23) show that Model 2 (χ^2 = 13.54, P = 0.0089**) and Model 4 (χ^2 = 39.584, P = 0.03216 *) are significantly better at capturing the data compared to the baseline Sociodemographic model. When comparing Model 2 and Model 4, the χ^2 results reveal that Model 4 does not improve the fit of Model 2. Additionally, Model 2 shows the best fit according to the AIC criterion, while the BIC score indicates that Model 1 provides the best fit, and the log-likelihood score prefers Model 4. Because the χ^2 and AIC point to Model 2, it can be concluded that this is probably the best fitting model for the mobile data.

Table 23: Mobile: HMMs comparisons with depression (PHQ-9 score) as the dependent variable. The first quadrant reports the one-sided χ^2 test result of Model 1 against the complex models (Model 2,3,4), including the added degrees of freedom and the χ^2 statistic. The second quadrant reports the one-sided χ^2 test results between Model 2 and Model 4.

	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1	11.00	13149.07	13212.41	-6563.54	13127.07			
Model 2	15.00	13143.53	13229.91	-6556.77	13113.53	13.54	4	0.0089**
Model 3	18.00	13152.06	13255.71	-6558.03	13116.06	11.015	7	0.138
Model 4	36.00	13159.49	13366.79	-6543.74	13087.49	39.584	25	0.03216 *
Model 2	15.00	13143.53	13229.91	-6556.77	13113.53			
Model 4	36.00	13159.49	13366.79	-6543.74	13087.49	26.04	21	0.2048

For the desktop data (Table G2), the more complete model (Model 4: Sociodemographic + Aggregate Volume + Temporal + Semantic) shows that age, income and the degree of urbanization have statistically significant negative associations with depression, while the number of days with tobacco use and the total duration spent on *message boards and forums* have statistically significant positive associations with depression.

According to the β coefficient, every extra minute spent on *message boards and forums* increases the PHQ-9 scores by 0.014 (95% CI (0.005–0.024), β_{std} = 0.051, P < 0.01).

The ICC value for all models is 0.77, indicating that the individual panelist random effect explain 77% of the variance explained that the model. The fixed effects explain 10% (marginal R^2) of the total variance explained by the most complex model, indicating that changes in the PHQ-9 scores are mostly heavily dependent on the individual panelist random effect. The one sided χ^2 test (Table 24) reveals that none of the more complex models (Model 2 to 4) are significantly better at capturing the

data than the baseline Sociodemographic model (Model 1). The baseline model also returns the better AIC and BIC results, although the more complex model (Model 4) has the best log-likelihood score.

Table 24: Desktop: HMMs comparisons with depression (PHQ-9 score) as dependent variable. The the one-sided χ^2 test result is reported for the test of Model 1 against the more complex models (Model 2,3,4), including the added degrees of freedom and the χ^2 statistic.

	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1	11.00	12826.21	12889.27	-6402.10	12804.21			
Model 2	13.00	12829.77	12904.30	-6401.88	12803.77	0.44	2	0.8033
Model 3	14.00	12830.68	12910.94	-6401.34	12802.68	1.5313	3	0.6751
Model 4	25.00	12837.78	12981.11	-6393.89	12787.78	16.424	14	0.2881

4.2.2 Suicide Risk

For each of the specified model presented in section 3.6, the model results with the PHQ-9-Q9 suicide risk score as the dependent variable for the mobile and desktop devices are shown in Tables G3–G4 in Appendix G respectively.

For the mobile data (Table G3), the more complete model (Model 4: Sociodemographic + Aggregate Volume + Temporal + Semantic) shows that age and income have statistically significant negative associations with suicide risk, while the total duration spent on *chat and messaging* platforms, the number of *health* related app views, and the number of *job-related* URL views have statistically significant positive associations with suicide risk.

Every extra minute spent on *chat and messaging* platforms increases the PHQ-9-Q9 suicide risk score by 0.0002 (95% CI (0.0001—0.0003), β_{std} = 0.061, P < 0.01), every extra *health* app view increases the PHQ-9-Q9 score by 0.0004 (95% CI 0.0000-–0.0008, β_{std} = 0.046, P < 0.05), and every extra *job-related* URL corresponds to an increase of 0.002 (95% CI 0.000-–0.004, β_{std} = 0.050, P < 0.05) in the PHQ-9-Q9 score. According to the standardized coefficients, the time spent on *chat and messaging* platforms has the highest impact on the suicide risk score among the statistically significant internet usage features.

The ICC value for all models was 0.63, indicating that individual panelist random effect explain 63% of the variance explained by the model. The fixed effects explain only 5% of the variance in the most complex model. The one-sided χ^2 test (Table 25) reveals that Model 4 (Sociodemographics + Aggregate Volume + Temporal + Semantic) is significantly better at capturing the data than the baseline Sociodemographic model. The AIC and the BIC score show that Model 1 has the better fit, whereas the log-likelihood value is the highest for Model 4. Weighting more heavily on the χ^2 test, Model 4 is the best fit.

For the desktop data (Table G4), the more complete model (Model 4: Sociodemographic + Aggregate Volume + Temporal + Semantic) shows that age and income have statistically significant negative associations with suicide risk, while the number of

Table 25: Mobile: HMMs comparisons with suicide risk (PHQ-9-Q9 score) as dependent variable. The the one-sided χ^2 test result is reported for the test of Model 1 against the more complex models (Model 2,3,4), including the added degrees of freedom and the χ^2 statistic.

	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1	11.00	3899.77	3963.11	-1938.88	3877.77			
Model 2	15.00	3901.68	3988.06	-1935.84	3871.68	6.09	4	0.1928
Model 3	18.00	3906.80	4010.45	-1935.40	3870.80	6.9681	7	0.4322
Model 4	36.00	3910.64	4117.94	-1919.32	3838.64	39.131	25	0.03574 *

days with tobacco use shows a statistically significant positive association with suicide risk. None of the internet usage features show significant associations.

The ICC value for all models was 0.64, indicating that individual panelist random effect explain 64% of the variance explained by the model. The fixed effects explain only 5% of the variance in the most complex model. The one sided χ^2 test (Table 26) reveals that none of the more complex models (Model 2 to 4) are significantly better at capturing the data than the baseline Sociodemographic model (Model 1). The AIC and the BIC score show that Model 1 has the better fit, whereas the log-likelihood value is the highest for Model 4. Overall, Model 1 is the best fit.

Table 26: Desktop: HMMs comparisons with suicide risk (PHQ-9-Q9 score) as dependent variable. The the one-sided χ^2 test result is reported for the test of Model 1 against the more complex models (Model 2,3,4), including the added degrees of freedom and the χ^2 statistic.

	npar	AIC	BIC	logLik	deviance	χ^2	Df	Pr(>χ^2)
Model 1	11.00	3662.08	3725.15	-1820.04	3640.08			
Model 2	13.00	3665.68	3740.22	-1819.84	3639.68	0.40	2	0.8186
Model 3	14.00	3662.42	3742.68	-1817.21	3634.42	5.6655	3	0.1291
Model 4	25.00	3682.91	3826.24	-1816.45	3632.91	7.175	14	0.9278

5 Discussion

The three objectives this work aimed to achieve have been addressed through the extensive pre-processing and feature creation, the classification analysis for depression and suicide risk classification, and the hierarchical models analysis for uncovering statistically significant internet usage features for depression and suicide risk. The following sections discuss in details how each objective has been targeted and what conclusions can be drawn from the conducted analyses. Additionally, possible limitations and future directions are considered.

5.1 Objective 1: To quantify internet usage (IU) from desktop and mobile traces in terms of volume, temporal and semantic features

Section 3.2 and section 3.3 have described how raw web browsing traces from desktop devices and raw web browsing and app usage traces from mobile devices have been pre-processed to create features indicative of the volume of internet usage in a PHQ-9 period. Time series of different granularity (URL, apps, sub-level-domains, on-off events, subcategories, parent categories, and others) have been created from the raw traces to interpolate different levels of data coarseness that can be useful to infer relevant user behaviours. Most of the known limitations of the data collection method have been addressed, including targeting timed-out URL views, duplicated views and inconsistent categorization across device types, to have a more realistic representation of user behaviour. The processed time series have then been used to create features representative of the total volume of internet usage (Aggregate Volume), the volume of internet usage by time of day and time of week (Temporal), the volume of internet usage by viewed content (Semantic), the volume of internet usage by viewed content in a specific time period (Semantic Temporal), and the randomness in user behaviours (Entropies and KL). The creation of the feature sets has been done by considering aspects related to depression that are backed up by psychological studies [17][12], and previous research on the associations between internet usage and depression and suicide risk (see Table 2.)

The main limitation pertaining the feature creation is in relation to the semantic features, which at the core are based on the Webshrinker [30] classification of domains for URL traces and a custom translation from app names to Webshrinker categories for app traces (Appendix C–D). The Webshrinker categorization of domains is not fully reliable as it is sometimes inaccurate. By consequence, the app re-categorization might also be biased, because it partially relies on string matching app names to categorized Webshrinker domains. Additionally, uncategorized apps and domains have not been included in the semantic features, possibly misrepresenting the behaviour of panelists who have a high presence of these domains and apps in their histories, such as people from less common ethnicities and languages.

5.2 Objective 2: To explore the potential of the created IU features for depression classification and suicide risk detection with ML models, and identify the best performing feature set

Section 3.5 explored the potential of the created feature sets for binary depression status classification and suicide risk detection, analyzing the performance changes from least privacy intrusive sets to more privacy intrusive sets. The classification analysis is performed separately for two feature selection methods: an exploratory analysis, where features are selected using recursive feature elimination with cross-validation

(RFECV), and a limited analysis, where features are pre-selected based on associations found in previous studies. For each selection method, two PHQ-9 depression splits are explored, the *Extremes* split to recognize people with no depression symptoms (PHQ-9 = 0) from people with high depression severity (PHQ-9 $\geq$ 15), and the *Minimal - Mild Up* split to recognize people with minimal depression severity (PHQ-9 < 5) from people with mild or greater depression severity (PHQ-9 $\geq$ 5). To assess suicide risk, the *No Risk - Suicide Risk* PHQ-9 question 9 split is explored to observe the potential of the created features set in recognizing people with no suicide risk symptoms (PHQ-9-Q9 = 0) from people with suicide risk symptoms (PHQ-9-Q9 > 0). In addition to observing the performance changes with more privacy intrusive IU sets, the performance is compared against the accuracy achieved from offline data (demographic and sociodemographic information), and the accuracy achieved with online and offline data (internet usage features sets in addition to demographic and sociodemographic information). The goal is to assess the standalone potential of internet usage data and its effectiveness when combined with or compared against sociodemographic information. The classification analyses are conducted using several classifiers and averaged across 15 train-test splitting seeds. Additionally, the most important features are observed using model specific feature importances or coefficients.

The results from the classification analysis reveal that sociodemographic information, specifically income and age information, detain the highest potential in classifying both depression and suicide risk, and that adding internet usage features from mobile or desktop devices does not improve the performance. The classification using IU feature sets shows that more privacy intrusive features do not always relate to better performance, and that aggregate volume information from internet usage is often the best performing feature set. The best performances with IU features (online) for desktop devices are 0.61 ± 0.04 (Aggregate Volume set), 0.56 ± 0.02 (Aggregate Volume set) and 0.54 ± 0.02 (Temporal set) for the *Extremes*, *Minimal - Mild Up* and *No Risk - Suicide Risk* splits respectively. For mobile devices, the best performance using IU features (online) is 0.59 ± 0.04 (Semantic Parent), 0.55 ± 0.02 (Aggregate Volume) and 0.52 ± 0.03 (Aggregate Volume) for the *Extremes*, *Minimal - Mild Up* and *No Risk - Suicide Risk* respectively. The performances improve significantly when adding offline information, with the best performances achieved with the Sociodemographic + All IU set for both mobile (*Extremes*: 0.65 BA, *Minimal - Mild Up*: 0.61, *No Risk - Suicide Risk*: 0.57 ± 0.03) and desktop devices (*Extremes*: 0.66 ± 0.05, *Minimal - Mild Up*: 0.60 ± 0.02, *No Risk - Suicide Risk*: 0.54 ± 0.02), but never outperform the accuracy achieved using offline features only from the *Sociodemographic* set for either desktop (Extremes: 0.72 ± 0.04, *Minimal - Mild Up*: 0.60 ± 0.02, *No Risk - Suicide Risk*: 0.63 ± 0.02) or mobile (*Extremes*: 0.73 ± 0.03, *Minimal - Mild Up*: 0.63 ± 0.02, *No Risk - Suicide Risk*: 0.59 ± 0.02). Among the used classifiers, there isn't one that always outperforms the others, although simpler classifiers such SVM-RBF and Logistic Regression appear to often return the best performance.

The lack of potential in the internet usage data, especially in the composite sets (All IU, Sociodemographic + All IU, and Demographic + All IU) when compared

to the individual sets, might be the results of poor model generalization. In the exploratory classification analysis, it is possible that this is the consequence of poor feature selection in the RFECV due to the large size of the composite features sets and increased model complexity. The validity of this argument diminishes in the context of a limited analysis, characterized by significantly smaller feature sets. To address this, further investigation, incorporating additional feature selection methods such as Recursive Feature Elimination (RFE) in the limited analysis, along with a detailed examination of training and testing performance, is needed to help determine whether the observed issues are attributable to poor model generalization or the confounding effects of certain included internet usage features. Comparison with similar papers [10][11] shows that the performances achieved in this study are less promising, which is speculated to be due to the lack of some high-potential features (number of calls sent and received, number of contacts saved), a larger and more heterogeneous populations, and differences in the definition of depressed individuals. Another key point to keep into consideration is the reliance of this study on self-reported measurements of depression from the PHQ-9 survey, which is self-administed and may suffer from biases, possibly affecting the classification results. Additionally, a more reliable and model-agnostic feature importance method should be used to contrast the results achieved from the best performing classifiers, for instance SHAP (SHapley Additive exPlanations) values [47], which would give an indication of both the strength and direction of the feature importances. SHAP values could also be incorporated into the feature selection process, such as in recursive feature elimination, as an alternative to relying on model-dependent coefficients, even though this approach can be computationally demanding.

The findings from the classification analysis reveal that, while there is potential in internet usage for depression and suicide risk detection, knowledge from sociodemographic information is the most useful factor in the classification. With regards to the level of privacy intrusiveness needed to achieve the best results from the explored features set from internet usage traces, low intrusiveness features, specifically features representing the aggregate volume of internet usage, often outperfom more privacy intrusive features, although the differences in classification accuracies are often minimal. Temporal semantic information from parent categories, and semantic information from interactivity categories have also been shown to have potential in depression classification.

In the light of sociodemographics factors being the best performing features in classifying depression, future analysis should focus on exploring the potential of the created IU features in depression classification for specific sub-groups, including individual age and income brackets, by employment status, and across genders. Future directions could also explore the classification performance on individual PHQ-9 questions, to investigate the potential of internet usage features in detecting the presence and severity of specific depression symptoms other than suicide risk, and how the detection of individual symptom can help with a better depression status assessment. There is an alternative approach to scoring the PHQ-9 survey for depression status

assessment [12], where an individual is considered depressed if it scores one or higher in specific questions of the questionnaire. This assessment method can be explored by first training the classifiers in recognizing individual symptoms from individual PHQ-9 questions, and using the outcome for each symptom to make a depression status assessment.

5.3 Objective 3: To identify which internet use measures correlate with depression and suicide risk when controlling for individual level characteristics and sociodemographic factors

Section 3.6 explored the associations between internet usage features and depression or suicide risk by using hierarchical mixed effects model with the features collected from the first three waves of the WebWell longitudinal study. The correlation analysis (section 3.4) showed statistically significant correlations between several internet usage features and depression or suicide risk, but the correlations fail to take into consideration potential confounding variables. In contrast, the hierarchical mixed-effects models provide a more robust approach by accounting for individual variations, capturing repeated measurements over time, and addressing potential confounding factors. This modeling strategy allows for a nuanced examination of the association between internet usage features and depression or suicide risk outcomes, offering a more comprehensive understanding that goes beyond mere correlations. To address individual level variations, the panelist identifier was used as a random effect, while sociodemographic, seasonal (wave number) and selected internet usage features are added as fixed effects for making inferences about associations existing in the general population. Four models definitions are explored for each dependent variable: a baseline model (Model 1) using sociodemographics fixed effects, a model using sociodemographic and aggregate volume internet usage fixed effects (Model 2), a model using sociodemographic, aggregate volume, and temporal fixed effects (Model 3), and a more complete model including sociodemographic, aggregate volume, temporal and semantic fixed effects (Model 4). The aim of the different model definitions is to observe statistically significant internet usage features, and explore which model suits the data the best. For each model, the considered features are pre-selected from existing literature and screened further using a VIF threshold of 1.5 to avoid multicollinearity and prevent overfitting. Model comparison is then done by observing the results of the one-sided χ^2 test with the baseline model (Model 1), and the AIC, BIC and log-likelihood values.

The results show that age and income have a negative effect on depression and suicide risk, while the number of days with tobacco use have a positive effect on depression and suicide risk (desktop population only). These findings are consistent with the findings in the existing literature [17][20]. Results from the most complete model (Model 4) show that, for desktop devices, the time spent on *message boards and forums* has a statistically significant positive association with depression (β =

0.014, 95% CI (0.005–0.024), β_{std}= 0.051, P < 0.01). For mobile devices, the count of app views during night time (β = 0.001, 95% CI (0.000–0.001), β_{std} = 0.042, P < 0.01), the total duration spent on *chat and messaging* platforms (β = 0.001, 95% CI (0.000–0.002), β_{std}= 0.046, P < 0.01) and the total count of *job-related* URL queries (β = 0.024, 95% CI (0.011–0.037), β_{std} = 0.075, P < 0.001) have statistically significant positive associations with depression. For suicide risk, the time spent on *chat and messaging* platforms (β = 0.0002, 95% CI (0.0001—0.0003), β_{std}= 0.061, P < 0.01), the number of *health* related apps (β = 0.0004, 95% CI 0.0000—0.0008, β_{std}= 0.046, P < 0.05) and the number of *job-related* URL visits (β = 0.002, 95% CI 0.000—0.004, β_{std} = 0.050, P < 0.05) have a statistically significant positive association with suicide risk severity for data from mobile devices.

The model comparison analysis show that both Model 4 and Model 2 fit the data better than the sociodemographic baseline (Model 1) for the depression analysis on mobile, and that Model 2 is better than Model 4. Model 2 shows that the average daily count of apps is a statistically significant positive predictor of depression, with every extra daily app view increasing the PHQ-9 score by 0.003 (95% CI (0.001–0.004), β_{std} = 0.066, P < 0.01). The average daily count of apps is not included in Model 3-4 because eliminated by the VIF threshold to avoid multicollinearity. Model 4 was the best model in the suicide risk analysis on mobile. For desktop data, Model 1 was the best fitting model in the depression analysis and in the suicide risk analysis.

For all models, the variance explained by the panelist random effect was the main contributor in explaining the total variance captured by the model, ranging from 0.64 to 0.77 depending on the device and dependent variable. Fixed effects, even in the baseline Sociodemographic models, are responsible for only a small portion of the variance explained, indicating that panelist individual characteristics explained more of the variability of the dependent variable than all of the fixed effects included. This is reasonable and to be expected, because the panelist random effect accounts for individual-specific characteristics that may not be fully captured by the fixed effects. Individual behaviors, preferences, or idiosyncrasies that contribute to the variability in the dependent variable are inherently better captured by the random effects associated with each panelist. This emphasizes the importance of considering individual-level variability, and it aligns with the understanding that not all sources of variation can be accounted for by general, population-level fixed effects, even when these include sociodemographic variables as done in all models. The substantial contribution of the panelist random effect highlights the significance of individual differences in explaining the observed variations in the explored models for depression and suicide risk.

The hierarchical model analysis reveals that there are statistically significant associations between internet usage features and depression or suicide risk, even when accounting for individual level characteristics and sociodemographic factors. Recognizing these associations can have crucial implications for developing targeted interventions and support strategies to mitigate suicide risk and depression severity.

The analysis on depression reveals that the daily count of app views, the count of app views in the night, the total time spent on *chat and messaging* platforms, the count of *job-related* URLs and the time spent of *message boards and forums* all showed statistically significant positive effect from either desktop or mobile data. It can be speculated that the significance of some of these features might be the online translation of several depressive symptoms more than the proof of a causal relationship with depression. For instance, the count of app view at night might be a symptom of sleep disturbances associated with depression (PHQ-9 question 3), and the number of *job-related* URLs might be a reflection of unhappiness at work and with ones' achievements (perhaps targeted by question 6 of the PHQ-9). It is challenging to attach meaning to the significance of certain variables, for instance the time spent on *message boards and forums*, without a more in-dept analysis of the viewed content. The analysis on suicide risk reveals that the time spent on *chat and messaging* platforms, the number of *health* related apps and the number of *job-related* URLs have a positive statistically significant association with suicide risk. These findings underline the importance of considering online communication patterns, health-related app usage, and exposure to job-related content as potential indicators for identifying individuals at an increased risk of suicidal behavior.

The main limitations of the hierarchical model analysis pertain the feature selection and the small number of observations per panelist. For the former, it might be appropriate to explore other feature selection methods for the fixed effects to include in the models. The current approach is based on a pre-selection of the semantic features from the existing literature, which might be non-comprehensive and outdated. Additionally, the VIF threshold used in this study was chosen arbitrarily to be very small to ensure negligible levels of multicollinearity between the variables, but it could be argued that a higher threshold or a smaller threshold would have been more appropriate. While the literature has proposed several VIF thresholds to detect multicollinearity, there is no general consensus and the choice is often left to be context dependent. The stringent VIF threshold was chosen in this case because several measurements (URLs, app, duration) of the same variable are likely to show collinearity. Regarding the sample size, the analyses from the hierarchical model should be expanded to include data points from all survey waves of the WebWell study. Currently, there are only three observations per panelist from the first three waves of the WebWell study, because the study was not complete yet at the time of this writing. Including the remaining osbervations from subsequent survey waves will enhance the statistical power of the analysis by improving the robustness of the results. A factor to consider is that sociodemographic features, specifically income, are considered static as they are taken only in the baseline survey, which is an assumption that likely holds true in the time frame considered in this analysis but might not hold when a longer time period is considered. For instance, changes in income that are not accounted for may affect the reliability of the statistical significance of certain features, such as job-related content viewing.

As per the classification analysis, future directions could explore whether there are statistically significant associations between internet usage and depression for specific

sub-groups, for instance by age bracket, which are not revealed in this population level analysis. Additionally, analyses on individual depression symptoms other than suicide risk could be conducted to find the internet features associated to specific depression markers. Lastly, to observe whether the statistical significance of the internet usage variables is due to a causal effect on the dependent variable, or due to a translation of other depressive symptoms into online behaviour, more psychological features should be included in the analysis. The could be an assessment from sleep quality and sleep disturbances from the Pittsburgh Sleep Quality index (PSQI), a loneliness score from the UCLA scale, and information about physical activity and diet, all of which are have been collected by the WebWell study.

6 Conclusion

This work aimed to assess the potential of web browsing and app usage traces from mobile or desktop devices for depression and suicide risk assessment, as well as finding associations between internet usage with depression and suicide risk. The implications of this study expand to early assessment of depression and suicide risk with data from desktop or mobile devices. The findings from the classification analysis reveal that, while there is potential in internet usage for depression and suicide risk detection, knowledge from sociodemographic information is the most useful factor in the classifications. Knowledge about the total volume of internet usage results to be a better predictor of depression status when compared to more privacy intrusive features, such as time-related quantity of internet usage features or semantic features of internet content, emphasizing that the relationship between utility and privacy intrusiveness in designing effective mental health monitoring systems is not always linear. Additionally, adding sociodemographic information to the internet usage features always improves performance, but never outperforms the results achieved using sociodemographics features alone. The consistent improvement in performance when incorporating sociodemographic information, coupled with its standalone efficacy, suggests the enduring relevance of traditional demographic factors in mental health evaluations. However, the study encourages a nuanced exploration of the potential of the internet usage features within diverse sub-groups, emphasizing the need for tailored approaches that consider age, income, substance use, and gender.

The hierarchical model analysis reveals that there are statistically significant associations between internet usage features and depression or suicide risk, even when accounting for individual level characteristics and sociodemographic factors. The analysis on depression reveals that the daily count of app views, the count of app views in the night, the total time spent on chat and messaging platforms, the time spent of message boards and forums and the number of job-related URLs all have statistically significant positive associations with depression PHQ-9 scores. The analysis on suicide risk reveals that the time spent on chat and messaging platforms, the number of health related apps and the number of job-related URLs have a positive statistically significant association with suicide risk PHQ-9-Q9 scores. The hierarchical model analysis

reinforces the existence of statistically significant associations between internet usage features and depression or suicide risk. The identified positive effects on depression and suicide risk from specific internet usage patterns highlight the potential for using online behaviors as markers for mental health conditions.

Collectively, these analyses advocate for a comprehensive and inclusive approach to mental health assessments that integrates both traditional sociodemographic factors and emerging internet usage patterns. The findings underline the need for sensitivity to privacy concerns while harnessing the potential of online behavioral data for mental health monitoring. Future research endeavors should delve into tailored analyses for specific demographic groups and explore associations with individual depressive symptoms, providing a more nuanced understanding of the complex interplay between internet usage and mental health.

References

[1] Liu Q, He H, Yang J, Feng X, Zhao F, Lyu J. Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study;126:134-40. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022395619307381.

[2] Organization WH. Depressive disorder; 2023. Accessed: January 21, 2024. Available from: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] Lotfi L, Flyckt L, Krakau I, Mårtensson B, Nilsson GH. Undetected depression in primary healthcare: occurrence, severity and co-morbidity in a two-stage procedure of opportunistic screening. Nordic journal of psychiatry. 2010;64(6):421-7.

[4] Smith KM, Renshaw PF, Bilello J. The diagnosis of depression: current and emerging methods. Comprehensive Psychiatry. 2013;54(1):1-6. Available from: https://www.sciencedirect.com/science/article/pii/S0010440X12001186.

[5] Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. Journal of Multidisciplinary Healthcare. 2016;9:211-7. PMID: 27217764. Available from: https://www.tandfonline.com/doi/abs/10.2147/JMDH.S104807.

[6] Torous J, Kiang MV, Lorme J, Onnela JP, et al. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR mental health. 2016;3(2):e5165.

[7] Maatoug R, Oudin A, Adrien V, Saudreau B, Bonnot O, Millet B, et al. Digital phenotype of mood disorders: A conceptual and critical review;13:895860.

Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.895860/full.

[8] Fox G, Clohessy T, van der Werff L, Rosati P, Lynn T. Exploring the competing influences of privacy concerns and positive beliefs on citizen acceptance of contact tracing mobile applications. Computers in Human Behavior. 2021;121:106806. Available from: https://www.sciencedirect.com/science/article/pii/S0747563221001291.

[9] Katikalapudi R, Chellappan S, Montgomery F, Wunsch D, Lutzen K. Associating Internet Usage with Depressive Behavior Among College Students;31(4):73-80. Available from: https://ieeexplore.ieee.org/document/6387969/.

[10] Yue C, Ware S, Morillo R, Lu J, Shang C, Bi J, et al. Automatic depression prediction using Internet traffic characteristics on smartphones;18:100137. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2352648320300295.

[11] Razavi R, Gharipour A, Gharipour M. Depression screening using mobile phone usage metadata: a machine learning approach;27(4):522-30. Available from: https://academic.oup.com/jamia/article/27/4/522/5715569.

[12] Kroenke K, Spitzer RL, Williams JBW. The PHQ-9. Journal of General Internal Medicine. 2001;16(9):606-13. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-1497.2001.016009606.x.

[13] Miranda Baxa AL Andrea Chipman, Puhl T. Depression in Europe: Building resilience through awareness, improved access, integrated care, and parity of esteem. Economist Impact; 2023. Accessed: December 12, 2023. Available from: https://impact.economist.com/projects/depression-in-europe/files/janssen-depression_in_europe_report.pdf.

[14] Eurostat. 7.2 Accessed: December 12, 2023. Available from: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210910-1.

[15] Santomauro DF, Herrera AMM, Shadid J, Zheng P, Ashbaugh C, Pigott DM, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. The Lancet. 2021;398(10312):1700-12.

[16] Owens MPJ David C, Davenport R. Clinical Assessment: Interviewing and Examination. In: Companion to Psychiatric Studies. eighth ed. St. Louis: Churchill Livingstone; 2010. p. 199-226.

[17] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Arlington, VA: American Psychiatric Association; 2013.

[18] Naslund JA, Aschbrenner KA, Araya R, Marsch LA, Unützer J, Patel V, et al. Digital Technology for Treating and Preventing Mental Disorders in Low-income and Middle-income Countries: A Narrative Review of the Literature. The Lancet Psychiatry. 2017 June;4(6):486-500.

[19] Shorey S, Ng ED, Wong CH. Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. British Journal of Clinical Psychology. 2022;61(2):287-305.

[20] Arias-de La Torre J, Vilagut G, Ronaldson A, Serrano-Blanco A, Martín V, Peters M, et al. Prevalence and variability of current depressive disorder in 27 European countries: a population-based study;6(10):e729-38. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2468266721000475.

[21] Organization WH. International Statistical Classification of Diseases and Related Health Problems, Eleventh Revision (ICD-11). World Health Organization; 2019.

[22] Hökby S, Hadlaczky G, Westerlund J, Wasserman D, Balazs J, Germanavicius A, et al. Are Mental Health Effects of Internet Use Attributable to the Web-Based Content or Perceived Consequences of Usage? A Longitudinal Study of European Adolescents;3(3):e31. Available from: http://mental.jmir.org/2016/3/e31/.

[23] Lam SSM, Jivraj S, Scholes S. Exploring the Relationship Between Internet Use and Mental Health Among Older Adults in England: Longitudinal Observational Study;22(7):e15683. Available from: https://www.jmir.org/2020/7/e15683.

[24] Nie D, Ning Y, Zhu T. Predicting Mental Health Status in the Context of Web Browsing. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. IEEE;. p. 185-9. Available from: http://ieeexplore.ieee.org/document/6511674/.

[25] Purwandari B, Wibawa WS, Fitriah N, Christia M, Bintari DR. Internet Addiction and Mental Health Prediction Using Ensemble Learning Based on Web Browsing History. In: Proceedings of the 3rd International Conference on Software Engineering and Information Management. ACM;. p. 155-9. Available from: https://dl.acm.org/doi/10.1145/3378936.3378947.

[26] Bessière K, Kiesler S, Kraut R, Boneva BS. EFFECTS OF INTERNET USE AND SOCIAL RESOURCES ON CHANGES IN DEPRESSION;11(1):47-70. Available from: http://www.tandfonline.com/doi/abs/10.1080/13691180701858851.

[27] Fan Zhang, Tingshao Zhu, Ang Li, Yilin Li, Xinguo Xu. A survey of web behavior and mental health. In: 2011 6th International Conference on Pervasive

Computing and Applications. IEEE;. p. 189-95. Available from: https://ieeexplore.ieee.org/document/6106503/.

[28] Müller A, Steins-Loeber S, Trotzke P, Vogel B, Georgiadou E, De Zwaan M. Online shopping in treatment-seeking patients with buying-shopping disorder. Comprehensive Psychiatry. 2019 Oct;94:152120. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010440X19300434.

[29] Berryman C, Ferguson CJ, Negy C. Social media use and mental health among young adults. Psychiatric quarterly. 2018;89:307-14.

[30] Webshrinker. Webshrinker Simplified Categories; 2023. Accessed: December 14, 2023. Available from: https://webshrinker.com/webshrinker-categories.

[31] Statista. Population of Germany as of December 31, 2022, by age group; 2022. Accessed: December 14, 2023. Available from: https://www.statista.com/statistics/454349/population-by-age-group-germany/.

[32] Statista. Average YouTube video length as of December 2018, by categoryp; 2018. Accessed: December 14, 2023. Available from: https://www.statista.com/statistics/1026923/youtube-video-category-average-length/.

[33] Aledavood T. Temporal patterns of human behavior [Doctoral thesis]. School of Science; 2017. Available from: http://urn.fi/URN:ISBN:978-952-60-7724-6.

[34] Aledavood T, Kivimäki I, Lehmann S, Saramäki J. Quantifying daily rhythms with non-negative matrix factorization applied to mobile phone data. Scientific reports. 2022;12(1):5544.

[35] Opoku Asare K, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study. JMIR Mhealth Uhealth. 2021 Jul;9(7):e26540. Available from: https://mhealth.jmir.org/2021/7/e26540.

[36] Aledavood T, Hoyos AMT, Alakörkkö T, Kaski K, Saramäki J, Isometsä E, et al. Data collection for mental health studies through digital platforms: requirements and design of a prototype. JMIR research protocols. 2017;6(6):e6919.

[37] Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of Mobile Apps for Depression and Anxiety: Cross-Sectional Survey Study on Patient Interest and Barriers to Engagement. JMIR Ment Health. 2019 Jan;6(1):e11334. Available from: http://mental.jmir.org/2019/1/e11334/.

[38] Ioannou A, Tussyadiah I. Privacy and surveillance attitudes during health crises: Acceptance of surveillance and privacy protection behaviours. Technology in Society. 2021;67:101774.

[39] Oliveira M, Yang J, Griffiths D, Bonnay D, Kulshrestha J. Browsing behavior exposes identities on the Web; 2023.

[40] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework; 2019.

[41] Contributors O. Optuna samplers; 2023. Accessed: December 27, 2023. https://optuna.readthedocs.io/en/stable/reference/samplers/index.html.

[42] Computing AS. Triton cluster; 2023. Accessed: December 27, 2023. https://scicomp.aalto.fi/triton/.

[43] Bates D, Mächler M, Bolker B, Walker S. lme4: Linear Mixed-Effects Models using 'Eigen' and S4; 2022. R package version 1.1-27. Available from: https://CRAN.R-project.org/package=lme4.

[44] Nakagawa S, Schielzeth H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution. 2013;4:133-42.

[45] Gelman A. Scaling regression inputs by dividing by two standard deviations. Statistics in medicine. 2008;27(15):2865-73.

[46] Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Quality & quantity. 2018;52:1957-76.

[47] Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions; 2017.

[48] Löwe B, Spitzer RL, Zipfel S, Herzog W. Gesundheitsfragebogen für Patienten (PHQ-D). Komplettversion und Kurzform. Testmappe mit Manual, Fragebögen, Schablonen. 2. Auflage. Pfizer GmbH; 2002.

[49] Gesiscss. Adding Category in Web Tracking Notebook;. Accessed: January 21, 2024. https://github.com/gesiscss/web_tracking/blob/master/research/web_routineness/release/00_1_Adding_Category.ipynb.

[50] Kulshrestha J, Oliveira M, Karaçalik O, Bonnay D, Wagner C. Web Routineness and Limits of Predictability: Investigating Demographic and Behavioral Differences Using Web Tracking Data. CoRR. 2020;abs/2012.15112. Available from: https://arxiv.org/abs/2012.15112.

A PHQ-9 questionnaire

Table A1: Patient Health Questionnaire (PHQ-9) [12] with the official german translation [48] used in the WebWell study. The PHQ-9 score is the sum of the questions scores. Question 9, referred in the main body as PHQ-9-Q9, is used to threshold panelist by suicide risk status.

English	**German**
Instructions	
Over the last 2 weeks, how often have you been bothered by any of the following problems?	*Wie oft fühlten Sie sich im Verlauf der letzten2 Wochen durch die folgenden Beschwerdenbeeinträchtigt?*
Questions	
1. Little interest or pleasure in doing things	1. Wenig Interesse oder Freude an Ihren Tätigkeiten
2. Feeling down, depressed, or hopeless	2. Niedergeschlagenheit, Schwermut oder Hoffnungslosigkeit.
3. Trouble falling or staying asleep, or sleeping too much	3. Schwierigkeiten ein- oder durchzuschlafen oder vermehrter Schlaf
4. Feeling tired or having little energy	4. Müdigkeit oder Gefühl, keine Energie zu haben
5. Poor appetite or overeating	5. Verminderter Appetit oder übermäßiges Bedürfnis zu essen
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	6. Schlechte Meinung von sich selbst; Gefühl, ein Versager zu sein oder die Familie enttäuscht zu haben
7. Trouble concentrating on things, such as reading the newspaper or watching television	7. Schwierigkeiten, sich auf etwas zu konzentrieren, z.B. beim Zeitunglesen oder Fernsehen
8. Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	8.Waren Ihre Bewegungen oder Ihre Spracheso verlangsamt, dass es auch anderenauffallen würde? Oder waren Sie im Gegenteil „zappelig“ oder ruhelos und hatten dadurcheinen stärkeren Bewegungsdrang als sonst?
9. Thoughts that you would be better off dead or of hurting yourself in some way	9. Gedanken, dass Sie lieber tot wären oder sich Leid zufügen möchten
Answer options	
0 - Not at all 1 - Several days 2 - More than half the days 3 - Nearly every day	0 - Überhauptnicht 1 - AneinzelnenTagen 2 - An mehrals derHälfte derTage 3 - Beinahejeden Tag

B Addition of sub-categories in URL traces

The URL traces from mobile and desktop devices have domain related Webshrinker categories as explained in section 3.2.1. Nevertheless, URL traces from very common google services, such as scholar.google.com and docs.google.com, would all be labelled with the Webshrinker category *search-engines and portals* because they share the same domain google.com. To overcome this limitation, sub level domains (SLD) are created from the URL field of the URL traces and common SLD domains are re-categorized with more appropriate categories. Sub level domains are in the form of subdomain.domain.top_level_domain extracted from the URL (scholar.google.com/digitalphenotyping... → scholar.google.com).
Known SLDs are used to add more informative categories to the URL views, for instance the URL trances with SLD scholar.google.com are re-categorized from *search-engines and portals* to *search-engines and portals, education* by adding the *education* category.
Additionally, the *email* category is added from a known list of domains providing email services individually labelled from a previous study on a similar dataset [49][50], and the *productivity* category is added for productivity related google services. A summary the conditions applied to the URL traces is shown in Table B1, many of which are taken directly from a previous study [49][50] on a similar dataset.

Table B1: Added categories for the URL traces on the basis of URL strings, domains or sub-level domains (SLD)

Sub-category	**Condition**
email	Domain in list of email domains [49] and SLD contains the string "mail"
productivity	SLD in: docs.google.com caldendar.google.com office.google.com photos.google.com drive.google.com onedrive.live.com
social-networking	SLD in: plus.google.com drive.yahoo.com groups.vodafone.de
news and media	SLD in: news.google.com magazin.vodafone.de x.enews.vodafone.de
translators	SLD in: translate.google.com translate.google.de
entertainment	URL contains string "amazon.de/gp/video"
education	SLD in: scholar.google.com
travel	URL contains strings "google.com/maps" or "google.de/maps". Or SLD in: flights.google.com flights.google.de
shopping	SLD in: play.google.com
survey	SLD contains the string "survey"

C App views sub-category string matching

As introduced in section 3.2.2, the second step of re-categorizing app views with the sub-category set is to string match the app name to know string indicative of a specific subcategories. If the string is present in the app name, the app view is re-categorized with the sub-category. Table C1 summarizes the list of strings used to string match the app name to the sub-category. The strings were manually collected by observing the top app views by count, and by logical association. A new sub-cateogory, *tools*, is created specifically to label default phone apps such as launchers, home, clock and hardware.

Table C1: Strings used in string matching app names to sub-categories. If the string is present in the app name, the sub-category is added to the app view.

Sub-category	**Strings to match in app names**
dating and personals	dating, meet, gay, queer
email	mail
health	health, fit, calorie, kalorie, step, fitness, counter, sleep, walk, band
productivity	reader, calendar, document, note, calculator, scan, editor, planner
games	game, play, puzzle, ™, ®
tools	launcher, starter, home, camera, record, mic, com.miui, com.samsung, com.android, com.google, settings, systems, phone, security, contact, files, galerij, gallery, camara, clock, klok, seguridad, dialer, software, update, photo, battery, video, control, monitor, authenticator, file, manager
education	learn, language, podcast, course
entertainment	music, mp3, youtube, radio, podcast, tv
news and media	news, weather, wetter
search-engines and portals	explorer, browser
economy and finance	finance, bank
chat and instant-messaging	chat, sms, mms, messenger, messag, text
survey	survey, mingle, panel
shopping	coupon, kauf, deal
travel	transport, route, train, map, auto, DB
food and recepies	food, essen, delivery, rezepte, recip, meal, takeway, grocer
message boards and forums	reddit

D App views app category to sub-category

The last step in labelling the app views with sub-category is to use the provided app category from the PlayStore to match to a sub-category, in the case where the app name is not found in the top 600 labelled apps or no match is found by string matching as explained in Appendix C. The app category to sub-category matching is shown in Table D1.

Table D1: Sub-category matching to app category from the PlayStore for categorization of apps which haven't matched to a domain or to a string as described in section 3.2.2 and Appendix C.

Sub-category	**App category from the PlayStore**
real-estate	Home & House
productivity	Art & Design
travel	Travel, transportation & navigation
tools	Tools, computers & electronics Pre-Installed Parenting
chat and messaging	Email, Messaging & Telephone
entertainment	Audio, video & entertainment
food and recepies	Lifestyle, food & nightlife
sport	Hobbies & sports
health	Health & fitness
business	Business & Industrial Lifestyle, food & nightlife Beauty
shopping	Shopping & price comparison
news and media	News, Media & Publications
education	Jobs & Education

E Exploratory classification results

Table E1: Mobile: Extremes (PHQ-9 = 0 - PHQ-9 ≥ 15) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.53 ± 0.04	0.53 ± 0.04	0.51 ± 0.04	0.49 ± 0.05	0.43 ± 0.06	0.45 ± 0.05	0.39 ± 0.09	0.36 ± 0.08	0.41 ± 0.08	0.47 ± 0.10	0.41 ± 0.13	0.36 ± 0.10	0.64 ± 0.08	0.59 ± 0.11	0.61 ± 0.12	0.62 ± 0.09
Temporal	0.52 ± 0.03	0.54 ± 0.03	0.55 ± 0.05	0.51 ± 0.03	0.42 ± 0.04	0.40 ± 0.06	0.45 ± 0.08	0.43 ± 0.05	0.40 ± 0.06	0.36 ± 0.08	0.44 ± 0.11	0.43 ± 0.06	0.65 ± 0.07	0.71 ± 0.08	0.65 ± 0.07	0.60 ± 0.05
Semantic Subcategories	0.53 ± 0.03	0.56 ± 0.03	0.54 ± 0.02	0.54 ± 0.03	0.42 ± 0.06	0.53 ± 0.04	0.43 ± 0.04	0.45 ± 0.04	0.40 ± 0.08	0.59 ± 0.08	0.39 ± 0.06	0.43 ± 0.05	0.66 ± 0.08	0.54 ± 0.10	0.69 ± 0.05	0.64 ± 0.04
Semantic Parent	0.52 ± 0.04	0.53 ± 0.03	0.53 ± 0.03	0.53 ± 0.05	0.43 ± 0.05	0.46 ± 0.06	0.41 ± 0.06	0.46 ± 0.06	0.41 ± 0.07	0.48 ± 0.07	0.37 ± 0.07	0.47 ± 0.08	0.64 ± 0.11	0.58 ± 0.07	0.70 ± 0.06	0.59 ± 0.08
Semantic Interactive	0.59 ± 0.04	0.55 ± 0.04	0.57 ± 0.03	0.55 ± 0.04	0.51 ± 0.07	0.49 ± 0.05	0.48 ± 0.03	0.48 ± 0.06	0.50 ± 0.08	0.52 ± 0.08	0.44 ± 0.05	0.49 ± 0.08	0.68 ± 0.07	0.58 ± 0.09	0.70 ± 0.07	0.62 ± 0.06
Entropies and KL	0.47 ± 0.03	0.50 ± 0.03	0.47 ± 0.03	0.51 ± 0.05	0.34 ± 0.06	0.41 ± 0.07	0.37 ± 0.06	0.42 ± 0.07	0.35 ± 0.10	0.44 ± 0.09	0.39 ± 0.10	0.42 ± 0.08	0.59 ± 0.12	0.56 ± 0.09	0.55 ± 0.12	0.60 ± 0.06
Temporal Semantic Interactive	0.57 ± 0.03	0.58 ± 0.03	0.57 ± 0.02	0.56 ± 0.04	0.46 ± 0.04	0.45 ± 0.05	0.42 ± 0.06	0.47 ± 0.05	0.41 ± 0.05	0.39 ± 0.07	0.37 ± 0.07	0.44 ± 0.08	0.73 ± 0.06	0.78 ± 0.05	0.76 ± 0.06	0.69 ± 0.07
Temporal Semantic Parent	0.54 ± 0.04	0.57 ± 0.02	0.56 ± 0.03	0.57 ± 0.03	0.43 ± 0.06	0.48 ± 0.04	0.42 ± 0.05	0.48 ± 0.04	0.40 ± 0.08	0.47 ± 0.07	0.37 ± 0.06	0.47 ± 0.06	0.68 ± 0.06	0.67 ± 0.06	0.74 ± 0.05	0.66 ± 0.05
All IU	0.54 ± 0.02	0.56 ± 0.04	0.51 ± 0.04	0.55 ± 0.03	0.46 ± 0.04	0.50 ± 0.05	0.42 ± 0.05	0.47 ± 0.05	0.46 ± 0.06	0.51 ± 0.07	0.41 ± 0.09	0.45 ± 0.07	0.62 ± 0.05	0.62 ± 0.06	0.60 ± 0.09	0.66 ± 0.07
Demographic	0.61 ± 0.04	0.62 ± 0.04	0.63 ± 0.02	0.64 ± 0.05	0.54 ± 0.07	0.57 ± 0.05	0.57 ± 0.04	0.58 ± 0.06	0.53 ± 0.08	0.60 ± 0.08	0.56 ± 0.05	0.59 ± 0.08	0.69 ± 0.04	0.65 ± 0.06	0.71 ± 0.03	0.69 ± 0.08
Sociodemographic	0.70 ± 0.03	0.70 ± 0.04	0.73 ± 0.03	0.68 ± 0.03	0.65 ± 0.05	0.66 ± 0.04	0.69 ± 0.03	0.63 ± 0.04	0.65 ± 0.08	0.69 ± 0.06	0.70 ± 0.06	0.61 ± 0.06	0.75 ± 0.04	0.70 ± 0.05	0.76 ± 0.04	0.75 ± 0.05
Demographic + All IU	0.58 ± 0.04	0.55 ± 0.04	0.56 ± 0.03	0.54 ± 0.03	0.51 ± 0.04	0.48 ± 0.06	0.49 ± 0.04	0.46 ± 0.04	0.52 ± 0.06	0.50 ± 0.09	0.48 ± 0.07	0.44 ± 0.06	0.63 ± 0.06	0.60 ± 0.06	0.65 ± 0.08	0.64 ± 0.06
Sociodemographic + All IU	0.64 ± 0.04	0.55 ± 0.03	0.62 ± 0.04	0.57 ± 0.04	0.58 ± 0.04	0.49 ± 0.04	0.56 ± 0.05	0.48 ± 0.06	0.57 ± 0.06	0.52 ± 0.06	0.56 ± 0.07	0.46 ± 0.07	0.71 ± 0.05	0.57 ± 0.05	0.68 ± 0.06	0.67 ± 0.06

Table E2: Desktop: Extremes (PHQ-9 = 0 - PHQ-9 ≥ 15) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.51 ± 0.04	0.57 ± 0.05	0.54 ± 0.05	0.54 ± 0.04	0.38 ± 0.05	0.51 ± 0.05	0.42 ± 0.07	0.42 ± 0.06	0.38 ± 0.07	0.56 ± 0.07	0.42 ± 0.09	0.41 ± 0.07	0.63 ± 0.08	0.59 ± 0.09	0.65 ± 0.06	0.68 ± 0.05
Temporal	0.54 ± 0.04	0.48 ± 0.05	0.54 ± 0.03	0.50 ± 0.05	0.44 ± 0.05	0.39 ± 0.05	0.43 ± 0.07	0.39 ± 0.07	0.45 ± 0.08	0.43 ± 0.07	0.45 ± 0.09	0.41 ± 0.08	0.62 ± 0.06	0.52 ± 0.09	0.63 ± 0.07	0.59 ± 0.04
Semantic Subcategories	0.53 ± 0.04	0.53 ± 0.03	0.49 ± 0.04	0.55 ± 0.04	0.40 ± 0.07	0.43 ± 0.05	0.33 ± 0.07	0.43 ± 0.05	0.40 ± 0.08	0.45 ± 0.07	0.32 ± 0.07	0.42 ± 0.08	0.65 ± 0.08	0.61 ± 0.07	0.65 ± 0.07	0.68 ± 0.06
Semantic Parent	0.54 ± 0.03	0.52 ± 0.05	0.52 ± 0.04	0.51 ± 0.04	0.46 ± 0.04	0.46 ± 0.06	0.44 ± 0.07	0.40 ± 0.06	0.52 ± 0.07	0.53 ± 0.09	0.50 ± 0.11	0.40 ± 0.07	0.55 ± 0.06	0.51 ± 0.07	0.53 ± 0.10	0.62 ± 0.07
Semantic Interactive	0.51 ± 0.03	0.46 ± 0.03	0.52 ± 0.03	0.53 ± 0.04	0.37 ± 0.06	0.35 ± 0.06	0.37 ± 0.07	0.40 ± 0.05	0.36 ± 0.08	0.38 ± 0.08	0.36 ± 0.09	0.38 ± 0.07	0.66 ± 0.08	0.54 ± 0.10	0.68 ± 0.09	0.67 ± 0.06
Entropies and KL	0.49 ± 0.03	0.47 ± 0.02	0.50 ± 0.04	0.47 ± 0.03	0.39 ± 0.06	0.38 ± 0.05	0.46 ± 0.05	0.32 ± 0.04	0.44 ± 0.08	0.42 ± 0.08	0.57 ± 0.08	0.30 ± 0.04	0.54 ± 0.09	0.52 ± 0.08	0.43 ± 0.08	0.65 ± 0.05
Temporal Semantic Interactive	0.52 ± 0.02	0.47 ± 0.04	0.50 ± 0.02	0.57 ± 0.05	0.30 ± 0.09	0.32 ± 0.09	0.22 ± 0.08	0.46 ± 0.06	0.28 ± 0.11	0.35 ± 0.11	0.20 ± 0.13	0.43 ± 0.07	0.77 ± 0.10	0.59 ± 0.12	0.79 ± 0.14	0.71 ± 0.06
Temporal Semantic Parent	0.53 ± 0.04	0.52 ± 0.04	0.52 ± 0.03	0.60 ± 0.03	0.37 ± 0.06	0.40 ± 0.06	0.35 ± 0.06	0.49 ± 0.04	0.32 ± 0.06	0.41 ± 0.08	0.34 ± 0.09	0.46 ± 0.05	0.74 ± 0.08	0.62 ± 0.08	0.70 ± 0.09	0.75 ± 0.05
All IU	0.55 ± 0.03	0.52 ± 0.06	0.54 ± 0.05	0.58 ± 0.04	0.44 ± 0.05	0.44 ± 0.07	0.45 ± 0.05	0.47 ± 0.05	0.44 ± 0.08	0.50 ± 0.09	0.48 ± 0.08	0.44 ± 0.05	0.66 ± 0.07	0.54 ± 0.07	0.60 ± 0.08	0.72 ± 0.04
Demographic	0.64 ± 0.05	0.63 ± 0.03	0.63 ± 0.04	0.61 ± 0.03	0.55 ± 0.07	0.56 ± 0.04	0.56 ± 0.04	0.53 ± 0.04	0.57 ± 0.09	0.62 ± 0.08	0.58 ± 0.05	0.54 ± 0.08	0.70 ± 0.06	0.64 ± 0.09	0.68 ± 0.05	0.69 ± 0.08
Sociodemographic	0.68 ± 0.05	0.72 ± 0.04	0.70 ± 0.05	0.71 ± 0.04	0.62 ± 0.06	0.67 ± 0.05	0.64 ± 0.06	0.66 ± 0.05	0.66 ± 0.07	0.75 ± 0.06	0.69 ± 0.08	0.71 ± 0.06	0.71 ± 0.04	0.70 ± 0.06	0.71 ± 0.04	0.71 ± 0.05
Demographic + All IU	0.62 ± 0.03	0.52 ± 0.05	0.58 ± 0.05	0.64 ± 0.04	0.55 ± 0.04	0.44 ± 0.08	0.51 ± 0.05	0.56 ± 0.05	0.59 ± 0.06	0.50 ± 0.10	0.55 ± 0.06	0.53 ± 0.05	0.66 ± 0.04	0.55 ± 0.06	0.61 ± 0.07	0.75 ± 0.05
Sociodemographic + All IU	0.63 ± 0.03	0.51 ± 0.04	0.60 ± 0.04	0.64 ± 0.03	0.56 ± 0.04	0.43 ± 0.06	0.52 ± 0.05	0.55 ± 0.05	0.60 ± 0.07	0.48 ± 0.09	0.56 ± 0.07	0.56 ± 0.07	0.67 ± 0.05	0.54 ± 0.06	0.64 ± 0.06	0.72 ± 0.06

Table E3: Mobile: Minimal-Mild Up (PHQ-9 < 5 - PHQ-9 ≥ 5) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.55 ± 0.02	0.54 ± 0.02	0.53 ± 0.02	0.54 ± 0.02	0.55 ± 0.02	0.50 ± 0.03	0.44 ± 0.05	0.56 ± 0.02	0.51 ± 0.04	0.43 ± 0.05	0.36 ± 0.06	0.53 ± 0.05	0.59 ± 0.05	0.65 ± 0.05	0.70 ± 0.06	0.56 ± 0.06
Temporal	0.50 ± 0.02	0.51 ± 0.02	0.52 ± 0.02	0.52 ± 0.02	0.49 ± 0.03	0.53 ± 0.04	0.45 ± 0.04	0.57 ± 0.03	0.45 ± 0.05	0.51 ± 0.06	0.38 ± 0.05	0.58 ± 0.05	0.56 ± 0.06	0.51 ± 0.05	0.66 ± 0.07	0.47 ± 0.05
Semantic Subcategories	0.53 ± 0.02	0.52 ± 0.01	0.52 ± 0.02	0.51 ± 0.02	0.53 ± 0.02	0.54 ± 0.01	0.49 ± 0.06	0.58 ± 0.02	0.48 ± 0.04	0.50 ± 0.03	0.44 ± 0.08	0.60 ± 0.04	0.58 ± 0.06	0.54 ± 0.04	0.61 ± 0.08	0.42 ± 0.05
Semantic Parent	0.54 ± 0.02	0.53 ± 0.02	0.53 ± 0.02	0.52 ± 0.02	0.55 ± 0.03	0.52 ± 0.02	0.49 ± 0.04	0.55 ± 0.02	0.52 ± 0.05	0.48 ± 0.04	0.42 ± 0.06	0.54 ± 0.05	0.55 ± 0.06	0.58 ± 0.05	0.63 ± 0.05	0.51 ± 0.07
Semantic Interactive	0.54 ± 0.01	0.55 ± 0.02	0.54 ± 0.01	0.53 ± 0.02	0.51 ± 0.02	0.55 ± 0.03	0.44 ± 0.04	0.56 ± 0.02	0.44 ± 0.03	0.50 ± 0.05	0.35 ± 0.05	0.55 ± 0.03	0.63 ± 0.03	0.59 ± 0.05	0.72 ± 0.05	0.51 ± 0.04
Entropies and KL	0.50 ± 0.02	0.50 ± 0.01	0.49 ± 0.02	0.50 ± 0.02	0.60 ± 0.04	0.51 ± 0.04	0.53 ± 0.06	0.56 ± 0.03	0.67 ± 0.09	0.49 ± 0.07	0.54 ± 0.09	0.57 ± 0.05	0.34 ± 0.11	0.50 ± 0.07	0.44 ± 0.09	0.43 ± 0.05
Temporal Semantic Interactive	0.52 ± 0.01	0.54 ± 0.02	0.54 ± 0.01	0.51 ± 0.02	0.47 ± 0.02	0.48 ± 0.04	0.38 ± 0.03	0.54 ± 0.03	0.40 ± 0.02	0.40 ± 0.06	0.28 ± 0.03	0.53 ± 0.06	0.65 ± 0.03	0.68 ± 0.07	0.81 ± 0.02	0.48 ± 0.06
Temporal Semantic Parent	0.52 ± 0.02	0.52 ± 0.01	0.52 ± 0.01	0.50 ± 0.02	0.44 ± 0.04	0.47 ± 0.05	0.38 ± 0.02	0.55 ± 0.03	0.36 ± 0.04	0.42 ± 0.06	0.28 ± 0.02	0.56 ± 0.05	0.68 ± 0.05	0.62 ± 0.06	0.76 ± 0.03	0.45 ± 0.06
All IU	0.52 ± 0.01	0.52 ± 0.02	0.52 ± 0.02	0.53 ± 0.03	0.53 ± 0.03	0.52 ± 0.04	0.49 ± 0.03	0.59 ± 0.03	0.51 ± 0.05	0.49 ± 0.08	0.43 ± 0.04	0.60 ± 0.04	0.54 ± 0.05	0.55 ± 0.09	0.61 ± 0.06	0.46 ± 0.03
Demographic	0.61 ± 0.02	0.60 ± 0.02	0.59 ± 0.02	0.59 ± 0.02	0.62 ± 0.02	0.66 ± 0.02	0.60 ± 0.03	0.62 ± 0.03	0.59 ± 0.04	0.69 ± 0.04	0.57 ± 0.05	0.59 ± 0.04	0.63 ± 0.04	0.51 ± 0.03	0.61 ± 0.04	0.60 ± 0.04
Sociodemographic	0.63 ± 0.02	0.62 ± 0.02	0.63 ± 0.02	0.60 ± 0.02	0.65 ± 0.02	0.65 ± 0.02	0.65 ± 0.02	0.63 ± 0.02	0.63 ± 0.02	0.63 ± 0.03	0.63 ± 0.02	0.62 ± 0.04	0.63 ± 0.03	0.62 ± 0.04	0.63 ± 0.03	0.58 ± 0.05
Demographic + All IU	0.57 ± 0.02	0.55 ± 0.02	0.57 ± 0.02	0.55 ± 0.03	0.59 ± 0.03	0.54 ± 0.09	0.59 ± 0.02	0.61 ± 0.03	0.56 ± 0.04	0.54 ± 0.11	0.56 ± 0.03	0.62 ± 0.05	0.58 ± 0.06	0.55 ± 0.12	0.57 ± 0.04	0.49 ± 0.05
Sociodemographic + All IU	0.60 ± 0.02	0.55 ± 0.02	0.59 ± 0.01	0.58 ± 0.07	0.63 ± 0.02	0.57 ± 0.03	0.61 ± 0.02	0.60 ± 0.01	0.60 ± 0.04	0.53 ± 0.03	0.59 ± 0.04	0.58 ± 0.09	0.61 ± 0.05	0.58 ± 0.03	0.59 ± 0.05	0.57 ± 0.22

Table E4: Desktop: Minimal-Mild Up (PHQ-9 < 5 - PHQ-9 ≥ 5) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.56 ± 0.02	0.55 ± 0.02	0.56 ± 0.02	0.53 ± 0.02	0.54 ± 0.02	0.53 ± 0.02	0.51 ± 0.02	0.56 ± 0.02	0.50 ± 0.03	0.49 ± 0.03	0.45 ± 0.02	0.57 ± 0.03	0.61 ± 0.04	0.61 ± 0.02	0.68 ± 0.02	0.48 ± 0.03
Temporal	0.55 ± 0.02	0.55 ± 0.01	0.54 ± 0.02	0.52 ± 0.02	0.57 ± 0.02	0.55 ± 0.02	0.54 ± 0.05	0.56 ± 0.02	0.56 ± 0.03	0.52 ± 0.03	0.53 ± 0.08	0.58 ± 0.03	0.54 ± 0.04	0.57 ± 0.05	0.55 ± 0.08	0.46 ± 0.03
Semantic Subcategories	0.53 ± 0.02	0.53 ± 0.02	0.51 ± 0.02	0.51 ± 0.01	0.59 ± 0.02	0.54 ± 0.02	0.56 ± 0.04	0.55 ± 0.02	0.64 ± 0.04	0.54 ± 0.04	0.59 ± 0.09	0.56 ± 0.03	0.41 ± 0.05	0.52 ± 0.04	0.44 ± 0.09	0.47 ± 0.02
Semantic Parent	0.53 ± 0.01	0.53 ± 0.02	0.53 ± 0.01	0.52 ± 0.02	0.59 ± 0.03	0.55 ± 0.03	0.61 ± 0.04	0.56 ± 0.03	0.64 ± 0.06	0.56 ± 0.04	0.71 ± 0.09	0.57 ± 0.04	0.42 ± 0.05	0.49 ± 0.04	0.34 ± 0.08	0.47 ± 0.03
Semantic Interactive	0.52 ± 0.01	0.54 ± 0.02	0.52 ± 0.02	0.50 ± 0.02	0.58 ± 0.03	0.53 ± 0.03	0.62 ± 0.04	0.53 ± 0.02	0.63 ± 0.06	0.50 ± 0.04	0.75 ± 0.08	0.53 ± 0.04	0.42 ± 0.06	0.57 ± 0.03	0.30 ± 0.07	0.47 ± 0.04
Entropies and KL	0.54 ± 0.01	0.54 ± 0.01	0.55 ± 0.02	0.51 ± 0.01	0.60 ± 0.03	0.58 ± 0.03	0.61 ± 0.02	0.54 ± 0.02	0.65 ± 0.06	0.61 ± 0.06	0.66 ± 0.04	0.55 ± 0.04	0.43 ± 0.04	0.46 ± 0.04	0.43 ± 0.03	0.47 ± 0.03
Temporal Semantic Interactive	0.50 ± 0.03	0.55 ± 0.02	0.50 ± 0.01	0.50 ± 0.02	0.58 ± 0.05	0.57 ± 0.03	0.60 ± 0.08	0.55 ± 0.02	0.68 ± 0.09	0.58 ± 0.06	0.77 ± 0.13	0.58 ± 0.04	0.32 ± 0.06	0.52 ± 0.07	0.23 ± 0.12	0.43 ± 0.04
Temporal Semantic Parent	0.52 ± 0.02	0.55 ± 0.02	0.52 ± 0.02	0.54 ± 0.01	0.59 ± 0.06	0.52 ± 0.08	0.62 ± 0.07	0.58 ± 0.02	0.68 ± 0.09	0.51 ± 0.08	0.77 ± 0.10	0.62 ± 0.04	0.36 ± 0.08	0.60 ± 0.07	0.28 ± 0.09	0.46 ± 0.03
All IU	0.55 ± 0.03	0.55 ± 0.02	0.53 ± 0.02	0.54 ± 0.02	0.58 ± 0.02	0.55 ± 0.02	0.57 ± 0.03	0.57 ± 0.03	0.58 ± 0.03	0.54 ± 0.03	0.60 ± 0.06	0.56 ± 0.04	0.52 ± 0.05	0.56 ± 0.03	0.45 ± 0.07	0.52 ± 0.04
Demographic	0.58 ± 0.02	0.57 ± 0.02	0.57 ± 0.02	0.59 ± 0.02	0.58 ± 0.02	0.53 ± 0.05	0.55 ± 0.02	0.61 ± 0.04	0.55 ± 0.03	0.48 ± 0.08	0.52 ± 0.04	0.62 ± 0.07	0.60 ± 0.04	0.66 ± 0.07	0.62 ± 0.04	0.56 ± 0.05
Sociodemographic	0.60 ± 0.02	0.59 ± 0.02	0.60 ± 0.02	0.60 ± 0.02	0.62 ± 0.02	0.59 ± 0.02	0.62 ± 0.03	0.60 ± 0.03	0.63 ± 0.04	0.57 ± 0.04	0.62 ± 0.05	0.59 ± 0.04	0.57 ± 0.04	0.61 ± 0.05	0.57 ± 0.04	0.61 ± 0.03
Demographic + All IU	0.57 ± 0.01	0.56 ± 0.02	0.57 ± 0.02	0.56 ± 0.03	0.59 ± 0.02	0.57 ± 0.02	0.58 ± 0.02	0.60 ± 0.04	0.60 ± 0.03	0.55 ± 0.05	0.56 ± 0.03	0.61 ± 0.06	0.54 ± 0.04	0.57 ± 0.07	0.57 ± 0.05	0.52 ± 0.04
Sociodemographic + All IU	0.60 ± 0.02	0.55 ± 0.02	0.58 ± 0.02	0.53 ± 0.06	0.64 ± 0.03	0.55 ± 0.03	0.60 ± 0.03	0.55 ± 0.10	0.65 ± 0.04	0.53 ± 0.05	0.60 ± 0.04	0.54 ± 0.14	0.55 ± 0.03	0.58 ± 0.06	0.56 ± 0.05	0.53 ± 0.12

Table E5: Mobile: *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.50 ± 0.02	0.48 ± 0.03	0.52 ± 0.03	0.50 ± 0.02	0.28 ± 0.03	0.28 ± 0.03	0.26 ± 0.07	0.21 ± 0.04	0.40 ± 0.06	0.45 ± 0.09	0.39 ± 0.12	0.22 ± 0.07	0.59 ± 0.03	0.51 ± 0.09	0.64 ± 0.10	0.77 ± 0.06
Temporal	0.48 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.50 ± 0.02	0.26 ± 0.03	0.29 ± 0.03	0.24 ± 0.04	0.18 ± 0.05	0.37 ± 0.07	0.45 ± 0.11	0.28 ± 0.06	0.16 ± 0.05	0.59 ± 0.05	0.56 ± 0.10	0.72 ± 0.06	0.84 ± 0.02
Semantic Subcategories	0.49 ± 0.02	0.50 ± 0.02	0.50 ± 0.02	0.49 ± 0.02	0.27 ± 0.04	0.25 ± 0.05	0.23 ± 0.05	0.21 ± 0.03	0.39 ± 0.10	0.35 ± 0.10	0.28 ± 0.08	0.22 ± 0.05	0.59 ± 0.09	0.66 ± 0.08	0.72 ± 0.06	0.76 ± 0.04
Semantic Parent	0.51 ± 0.02	0.50 ± 0.03	0.51 ± 0.03	0.49 ± 0.03	0.28 ± 0.05	0.27 ± 0.05	0.25 ± 0.05	0.21 ± 0.04	0.39 ± 0.08	0.38 ± 0.09	0.32 ± 0.08	0.23 ± 0.05	0.62 ± 0.06	0.63 ± 0.04	0.69 ± 0.07	0.75 ± 0.04
Semantic Interactive	0.49 ± 0.02	0.51 ± 0.03	0.51 ± 0.02	0.50 ± 0.02	0.26 ± 0.03	0.27 ± 0.04	0.24 ± 0.03	0.23 ± 0.03	0.35 ± 0.06	0.38 ± 0.09	0.28 ± 0.08	0.25 ± 0.04	0.62 ± 0.04	0.63 ± 0.07	0.73 ± 0.08	0.75 ± 0.04
Entropies and KL	0.48 ± 0.02	0.48 ± 0.02	0.49 ± 0.01	0.52 ± 0.03	0.26 ± 0.05	0.22 ± 0.07	0.26 ± 0.05	0.23 ± 0.05	0.40 ± 0.07	0.38 ± 0.14	0.42 ± 0.13	0.24 ± 0.06	0.56 ± 0.07	0.57 ± 0.14	0.55 ± 0.13	0.80 ± 0.05
Temporal Semantic Interactive	0.51 ± 0.02	0.52 ± 0.02	0.51 ± 0.02	0.49 ± 0.02	0.28 ± 0.02	0.28 ± 0.03	0.24 ± 0.03	0.18 ± 0.04	0.35 ± 0.04	0.37 ± 0.08	0.23 ± 0.03	0.18 ± 0.05	0.67 ± 0.02	0.66 ± 0.06	0.79 ± 0.02	0.80 ± 0.05
Temporal Semantic Parent	0.51 ± 0.02	0.50 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.27 ± 0.05	0.24 ± 0.05	0.23 ± 0.03	0.24 ± 0.03	0.37 ± 0.09	0.31 ± 0.07	0.24 ± 0.07	0.28 ± 0.05	0.66 ± 0.07	0.68 ± 0.08	0.78 ± 0.06	0.72 ± 0.05
All IU	0.51 ± 0.02	0.47 ± 0.02	0.50 ± 0.02	0.49 ± 0.02	0.29 ± 0.03	0.21 ± 0.04	0.26 ± 0.05	0.21 ± 0.04	0.40 ± 0.11	0.28 ± 0.08	0.34 ± 0.10	0.22 ± 0.05	0.63 ± 0.10	0.66 ± 0.08	0.67 ± 0.10	0.76 ± 0.04
Demographic	0.57 ± 0.03	0.58 ± 0.03	0.57 ± 0.03	0.54 ± 0.01	0.37 ± 0.03	0.37 ± 0.03	0.37 ± 0.03	0.32 ± 0.03	0.57 ± 0.05	0.51 ± 0.05	0.55 ± 0.05	0.43 ± 0.07	0.56 ± 0.03	0.65 ± 0.04	0.58 ± 0.04	0.65 ± 0.06
Sociodemographic	0.58 ± 0.02	0.57 ± 0.03	0.58 ± 0.03	0.58 ± 0.03	0.38 ± 0.02	0.37 ± 0.03	0.38 ± 0.03	0.37 ± 0.03	0.57 ± 0.05	0.54 ± 0.06	0.58 ± 0.05	0.50 ± 0.07	0.59 ± 0.03	0.60 ± 0.05	0.58 ± 0.02	0.66 ± 0.05
Demographic + All IU	0.53 ± 0.02	0.51 ± 0.03	0.51 ± 0.03	0.52 ± 0.03	0.31 ± 0.03	0.29 ± 0.04	0.28 ± 0.05	0.26 ± 0.05	0.41 ± 0.06	0.39 ± 0.06	0.41 ± 0.12	0.28 ± 0.06	0.66 ± 0.06	0.64 ± 0.04	0.61 ± 0.09	0.76 ± 0.06
Sociodemographic + All IU	0.54 ± 0.02	0.50 ± 0.02	0.53 ± 0.02	0.53 ± nan	0.31 ± 0.04	0.25 ± 0.05	0.31 ± 0.04	0.30 ± nan	0.41 ± 0.07	0.33 ± 0.07	0.42 ± 0.07	0.33 ± nan	0.67 ± 0.06	0.67 ± 0.07	0.64 ± 0.05	0.74 ± nan

Table E6: Desktop: *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0) exploratory classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy				F1				Recall				Specificity			
Model	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB	LR	RF	SVML	XGB
Feature name																
Aggregate Volume	0.48 ± 0.02	0.49 ± 0.01	0.48 ± 0.01	0.50 ± 0.02	0.15 ± 0.07	0.24 ± 0.05	0.19 ± 0.07	0.20 ± 0.05	0.25 ± 0.12	0.40 ± 0.10	0.36 ± 0.17	0.24 ± 0.08	0.72 ± 0.13	0.59 ± 0.10	0.61 ± 0.18	0.76 ± 0.08
Temporal	0.54 ± 0.02	0.52 ± 0.02	0.53 ± 0.02	0.51 ± 0.02	0.31 ± 0.03	0.31 ± 0.02	0.29 ± 0.02	0.25 ± 0.03	0.46 ± 0.06	0.54 ± 0.08	0.40 ± 0.06	0.33 ± 0.07	0.62 ± 0.05	0.51 ± 0.08	0.66 ± 0.06	0.69 ± 0.06
Semantic Subcategories	0.48 ± 0.02	0.52 ± 0.03	0.49 ± 0.02	0.49 ± 0.02	0.20 ± 0.05	0.28 ± 0.03	0.22 ± 0.05	0.18 ± 0.04	0.33 ± 0.11	0.44 ± 0.07	0.38 ± 0.14	0.19 ± 0.05	0.62 ± 0.12	0.59 ± 0.04	0.60 ± 0.15	0.80 ± 0.04
Semantic Parent	0.51 ± 0.03	0.51 ± 0.03	0.49 ± 0.02	0.50 ± 0.02	0.30 ± 0.03	0.27 ± 0.05	0.26 ± 0.03	0.20 ± 0.04	0.56 ± 0.09	0.46 ± 0.11	0.46 ± 0.12	0.22 ± 0.09	0.46 ± 0.07	0.56 ± 0.11	0.52 ± 0.13	0.79 ± 0.11
Semantic Interactive	0.47 ± 0.02	0.48 ± 0.02	0.47 ± 0.02	0.50 ± 0.02	0.21 ± 0.05	0.21 ± 0.06	0.22 ± 0.05	0.19 ± 0.05	0.36 ± 0.10	0.39 ± 0.16	0.41 ± 0.14	0.20 ± 0.05	0.58 ± 0.12	0.56 ± 0.16	0.53 ± 0.15	0.81 ± 0.04
Entropies and KL	0.50 ± 0.02	0.48 ± 0.03	0.50 ± 0.03	0.49 ± 0.02	0.28 ± 0.02	0.25 ± 0.03	0.29 ± 0.03	0.18 ± 0.05	0.49 ± 0.04	0.44 ± 0.11	0.56 ± 0.08	0.23 ± 0.08	0.52 ± 0.03	0.51 ± 0.09	0.44 ± 0.07	0.75 ± 0.07
Temporal Semantic Interactive	0.50 ± 0.03	0.49 ± 0.03	0.52 ± 0.02	0.51 ± 0.02	0.27 ± 0.03	0.27 ± 0.03	0.24 ± 0.04	0.25 ± 0.03	0.43 ± 0.10	0.44 ± 0.06	0.30 ± 0.12	0.35 ± 0.08	0.58 ± 0.08	0.54 ± 0.04	0.75 ± 0.11	0.66 ± 0.06
Temporal Semantic Parent	0.50 ± 0.02	0.50 ± 0.03	0.48 ± 0.03	0.51 ± 0.02	0.25 ± 0.03	0.24 ± 0.05	0.22 ± 0.04	0.20 ± 0.03	0.34 ± 0.07	0.37 ± 0.11	0.32 ± 0.11	0.19 ± 0.04	0.66 ± 0.06	0.63 ± 0.09	0.63 ± 0.13	0.82 ± 0.04
All IU	0.50 ± 0.02	0.52 ± 0.02	0.49 ± 0.01	0.50 ± 0.02	0.27 ± 0.03	0.30 ± 0.02	0.22 ± 0.04	0.21 ± 0.03	0.40 ± 0.06	0.49 ± 0.05	0.32 ± 0.11	0.21 ± 0.05	0.60 ± 0.05	0.55 ± 0.04	0.65 ± 0.12	0.80 ± 0.04
Demographic	0.57 ± 0.01	0.59 ± 0.02	0.58 ± 0.02	0.57 ± 0.02	0.35 ± 0.01	0.36 ± 0.02	0.36 ± 0.02	0.34 ± 0.02	0.56 ± 0.03	0.49 ± 0.03	0.56 ± 0.03	0.45 ± 0.03	0.58 ± 0.02	0.69 ± 0.02	0.60 ± 0.02	0.69 ± 0.04
Sociodemographic	0.62 ± 0.02	0.60 ± 0.02	0.62 ± 0.02	0.63 ± 0.02	0.39 ± 0.02	0.37 ± 0.02	0.39 ± 0.02	0.41 ± 0.03	0.61 ± 0.03	0.54 ± 0.05	0.60 ± 0.03	0.60 ± 0.04	0.63 ± 0.02	0.65 ± 0.04	0.64 ± 0.02	0.67 ± 0.03
Demographic + All IU	0.55 ± 0.02	0.53 ± 0.03	0.55 ± 0.03	0.52 ± 0.02	0.32 ± 0.03	0.31 ± 0.03	0.30 ± 0.05	0.23 ± 0.03	0.47 ± 0.05	0.51 ± 0.05	0.44 ± 0.09	0.24 ± 0.06	0.63 ± 0.02	0.56 ± 0.04	0.66 ± 0.07	0.79 ± 0.06
Sociodemographic + All IU	0.56 ± 0.02	0.53 ± 0.03	0.56 ± 0.02	0.55 ± 0.02	0.33 ± 0.03	0.31 ± 0.03	0.32 ± 0.03	0.29 ± 0.05	0.48 ± 0.05	0.46 ± 0.04	0.45 ± 0.05	0.39 ± 0.08	0.64 ± 0.02	0.61 ± 0.05	0.67 ± 0.04	0.70 ± 0.06

F Limited classification results

Table F1: Mobile: Extremes (PHQ-9 = 0 - PHQ-9 ≥ 15) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.54 ± 0.02	0.57 ± 0.03	0.44 ± 0.06	0.45 ± 0.06	0.44 ± 0.10	0.41 ± 0.08	0.64 ± 0.08	0.73 ± 0.06
Temporal	0.52 ± 0.03	0.51 ± 0.02	0.34 ± 0.10	0.38 ± 0.03	0.33 ± 0.12	0.34 ± 0.05	0.70 ± 0.12	0.68 ± 0.07
Semantic	0.57 ± 0.03	0.55 ± 0.02	0.52 ± 0.03	0.45 ± 0.04	0.55 ± 0.05	0.41 ± 0.05	0.59 ± 0.05	0.69 ± 0.05
All IU	0.54 ± 0.04	0.54 ± 0.03	0.44 ± 0.06	0.43 ± 0.05	0.44 ± 0.09	0.39 ± 0.07	0.64 ± 0.10	0.70 ± 0.05
Demographic	0.64 ± 0.03	0.63 ± 0.02	0.58 ± 0.07	0.57 ± 0.03	0.60 ± 0.10	0.57 ± 0.06	0.68 ± 0.06	0.70 ± 0.05
Sociodemographic	0.70 ± 0.02	0.70 ± 0.03	0.65 ± 0.03	0.65 ± 0.04	0.66 ± 0.05	0.66 ± 0.07	0.73 ± 0.04	0.75 ± 0.03
Demographic + All IU	0.57 ± 0.03	0.60 ± 0.03	0.49 ± 0.05	0.51 ± 0.05	0.48 ± 0.08	0.47 ± 0.07	0.66 ± 0.08	0.74 ± 0.07
Sociodemographic + All IU	0.60 ± 0.03	0.65 ± 0.03	0.52 ± 0.04	0.58 ± 0.07	0.50 ± 0.06	0.56 ± 0.09	0.70 ± 0.09	0.74 ± 0.04

Table F2: Desktop: Extremes (PHQ-9 = 0 - PHQ-9 ≥ 15) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.50 ± 0.04	0.61 ± 0.04	0.32 ± 0.11	0.52 ± 0.05	0.37 ± 0.15	0.53 ± 0.07	0.63 ± 0.15	0.69 ± 0.05
Temporal	0.51 ± 0.05	0.52 ± 0.03	0.42 ± 0.07	0.33 ± 0.06	0.49 ± 0.11	0.29 ± 0.07	0.52 ± 0.09	0.75 ± 0.07
Semantic	0.51 ± 0.03	0.50 ± 0.03	0.42 ± 0.05	0.25 ± 0.07	0.49 ± 0.12	0.22 ± 0.10	0.53 ± 0.11	0.77 ± 0.11
All IU	0.51 ± 0.04	0.58 ± 0.04	0.38 ± 0.08	0.45 ± 0.06	0.40 ± 0.12	0.42 ± 0.08	0.61 ± 0.12	0.73 ± 0.06
Demographic	0.63 ± 0.04	0.65 ± 0.04	0.55 ± 0.06	0.57 ± 0.05	0.66 ± 0.13	0.61 ± 0.08	0.60 ± 0.10	0.68 ± 0.06
Sociodemographic	0.71 ± 0.05	0.63 ± 0.04	0.66 ± 0.05	0.56 ± 0.05	0.72 ± 0.08	0.58 ± 0.06	0.70 ± 0.06	0.69 ± 0.04
Demographic + All IU	0.62 ± 0.05	0.64 ± 0.04	0.53 ± 0.06	0.55 ± 0.05	0.55 ± 0.08	0.57 ± 0.08	0.68 ± 0.06	0.70 ± 0.06
Sociodemographic + All IU	0.66 ± 0.05	0.64 ± 0.03	0.59 ± 0.06	0.58 ± 0.04	0.62 ± 0.07	0.61 ± 0.06	0.70 ± 0.06	0.67 ± 0.05

Table F3: Mobile: Minimal-Mild Up (PHQ-9 < 5 - PHQ-9 ≥ 5) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.54 ± 0.02	0.54 ± 0.02	0.51 ± 0.02	0.50 ± 0.02	0.44 ± 0.03	0.42 ± 0.03	0.65 ± 0.04	0.66 ± 0.03
Temporal	0.51 ± 0.02	0.52 ± 0.01	0.50 ± 0.04	0.40 ± 0.02	0.46 ± 0.07	0.30 ± 0.02	0.56 ± 0.07	0.74 ± 0.03
Semantic	0.54 ± 0.01	0.54 ± 0.01	0.55 ± 0.02	0.49 ± 0.03	0.51 ± 0.03	0.42 ± 0.07	0.57 ± 0.03	0.66 ± 0.08
All IU	0.52 ± 0.02	0.54 ± 0.01	0.51 ± 0.03	0.53 ± 0.04	0.46 ± 0.04	0.48 ± 0.06	0.58 ± 0.03	0.60 ± 0.06
Demographic	0.60 ± 0.02	0.60 ± 0.02	0.66 ± 0.02	0.65 ± 0.02	0.68 ± 0.03	0.67 ± 0.03	0.52 ± 0.04	0.54 ± 0.03
Sociodemographic	0.62 ± 0.02	0.61 ± 0.02	0.65 ± 0.02	0.64 ± 0.02	0.63 ± 0.03	0.63 ± 0.03	0.61 ± 0.04	0.59 ± 0.04
Demographic + All IU	0.57 ± 0.02	0.60 ± 0.02	0.59 ± 0.02	0.63 ± 0.02	0.56 ± 0.03	0.61 ± 0.05	0.58 ± 0.03	0.59 ± 0.06
Sociodemographic + All IU	0.57 ± 0.02	0.61 ± 0.01	0.59 ± 0.02	0.62 ± 0.02	0.55 ± 0.03	0.58 ± 0.03	0.59 ± 0.04	0.65 ± 0.03

Table F4: Desktop: Minimal-Mild Up (PHQ-9 < 5 - PHQ-9 ≥ 5) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.54 ± 0.02	0.56 ± 0.02	0.54 ± 0.03	0.54 ± 0.02	0.52 ± 0.05	0.48 ± 0.02	0.56 ± 0.04	0.64 ± 0.03
Temporal	0.55 ± 0.02	0.54 ± 0.01	0.57 ± 0.03	0.64 ± 0.02	0.56 ± 0.05	0.76 ± 0.04	0.54 ± 0.06	0.32 ± 0.04
Semantic	0.51 ± 0.02	0.53 ± 0.01	0.56 ± 0.03	0.63 ± 0.02	0.60 ± 0.07	0.76 ± 0.05	0.42 ± 0.07	0.30 ± 0.04
All IU	0.55 ± 0.01	0.55 ± 0.01	0.52 ± 0.08	0.58 ± 0.02	0.50 ± 0.08	0.59 ± 0.05	0.59 ± 0.06	0.51 ± 0.06
Demographic	0.57 ± 0.02	0.59 ± 0.01	0.57 ± 0.04	0.61 ± 0.02	0.55 ± 0.07	0.62 ± 0.04	0.59 ± 0.06	0.55 ± 0.03
Sociodemographic	0.60 ± 0.02	0.58 ± 0.01	0.60 ± 0.03	0.59 ± 0.02	0.58 ± 0.05	0.57 ± 0.03	0.62 ± 0.04	0.59 ± 0.04
Demographic + All IU	0.57 ± 0.02	0.56 ± 0.02	0.57 ± 0.02	0.61 ± 0.02	0.55 ± 0.03	0.64 ± 0.04	0.58 ± 0.03	0.48 ± 0.05
Sociodemographic + All IU	0.56 ± 0.03	0.59 ± 0.02	0.58 ± 0.03	0.61 ± 0.02	0.59 ± 0.05	0.62 ± 0.03	0.53 ± 0.06	0.57 ± 0.03

Table F5: Mobile: *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.48 ± 0.02	0.49 ± 0.01	0.26 ± 0.06	0.12 ± 0.07	0.51 ± 0.18	0.15 ± 0.08	0.46 ± 0.18	0.84 ± 0.09
Temporal	0.49 ± 0.02	0.49 ± 0.01	0.17 ± 0.08	0.14 ± 0.05	0.35 ± 0.21	0.14 ± 0.06	0.62 ± 0.20	0.85 ± 0.06
Semantic	0.48 ± 0.02	0.49 ± 0.02	0.25 ± 0.05	0.14 ± 0.05	0.38 ± 0.12	0.14 ± 0.06	0.58 ± 0.12	0.85 ± 0.06
All IU	0.48 ± 0.01	0.50 ± 0.01	0.18 ± 0.07	0.09 ± 0.04	0.29 ± 0.15	0.08 ± 0.04	0.67 ± 0.15	0.91 ± 0.06
Demographic	0.58 ± 0.03	0.57 ± 0.03	0.37 ± 0.03	0.36 ± 0.03	0.50 ± 0.05	0.49 ± 0.05	0.65 ± 0.05	0.65 ± 0.03
Sociodemographic	0.58 ± 0.02	0.59 ± 0.02	0.38 ± 0.02	0.38 ± 0.03	0.57 ± 0.06	0.52 ± 0.05	0.59 ± 0.04	0.66 ± 0.03
Demographic + All IU	0.50 ± 0.02	0.52 ± 0.01	0.28 ± 0.03	0.25 ± 0.06	0.39 ± 0.06	0.33 ± 0.09	0.62 ± 0.04	0.70 ± 0.09
Sociodemographic + All IU	0.53 ± 0.02	0.52 ± 0.02	0.31 ± 0.02	0.28 ± 0.04	0.43 ± 0.05	0.38 ± 0.08	0.63 ± 0.03	0.66 ± 0.06

Table F6: Desktop: *No Risk - Suicide Risk* (PHQ-9-Q9 = 0 - PHQ-9-Q9 > 0) limited classification results. Results are averaged across 15 train-test split seeds and t-statistic 95% CI are reported for each metric. Some feature subsets may have been trained on less than 15 seeds, in which case the appriate number of degrees of freedoms is used in the CI calculation.

	Balanced Accuracy		F1		Recall		Specificity	
Model	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG	RF	SVMRBG
Feature name								
Aggregate Volume	0.49 ± 0.02	0.50 ± 0.02	0.22 ± 0.07	0.19 ± 0.07	0.40 ± 0.14	0.29 ± 0.11	0.59 ± 0.14	0.70 ± 0.11
Temporal	0.48 ± 0.02	0.49 ± 0.02	0.24 ± 0.05	0.18 ± 0.06	0.45 ± 0.14	0.23 ± 0.08	0.51 ± 0.13	0.76 ± 0.10
Semantic	0.48 ± 0.01	0.49 ± 0.01	0.17 ± 0.07	0.03 ± 0.04	0.36 ± 0.20	0.03 ± 0.03	0.60 ± 0.20	0.95 ± 0.04
All IU	0.49 ± 0.02	0.50 ± 0.01	0.23 ± 0.05	0.17 ± 0.06	0.36 ± 0.09	0.23 ± 0.10	0.63 ± 0.09	0.77 ± 0.11
Demographic	0.59 ± 0.02	0.56 ± 0.02	0.36 ± 0.02	0.34 ± 0.02	0.49 ± 0.03	0.52 ± 0.05	0.70 ± 0.03	0.60 ± 0.07
Sociodemographic	0.61 ± 0.02	0.58 ± 0.02	0.39 ± 0.02	0.36 ± 0.02	0.56 ± 0.03	0.54 ± 0.03	0.66 ± 0.04	0.62 ± 0.02
Demographic + All IU	0.55 ± 0.02	0.53 ± 0.03	0.32 ± 0.03	0.30 ± 0.03	0.46 ± 0.05	0.43 ± 0.05	0.64 ± 0.03	0.64 ± 0.06
Sociodemographic + All IU	0.56 ± 0.03	0.57 ± 0.03	0.33 ± 0.04	0.34 ± 0.03	0.45 ± 0.06	0.50 ± 0.05	0.68 ± 0.02	0.65 ± 0.03

G Hierarchical Mixed Effect Models Results

This section present the table results for the hierarchical mixed effect model analysis as presented in section 3.6. Tables G1–G2 report the results for the analysis with depression PHQ-9 score as the dependent variable for mobile and desktop devices respectively. Tables G3–G4 report the results for the analysis with suicide risk PHQ-9-Q9 score as the dependent variable for mobile and desktop devices respectively.

Table G1: Mobile: depression score (PHQ-9) hierarchical models results. For each model, the pre-selected features reported in Table 18 are further screened using a VIF threshold (VIF ≤ 1.5). The analysis is run using the features and survey responses from the first three waves of the WebWell longitudinal study. The standardized β coefficients (st.Beta) are calculated using the Gelman approach [45].

	Sociodemographics				Sociodemographics + Aggregate Volume				Sociodemographics + Aggregate Volume + Temporal				Sociodemographics + Aggregate Volume + Temporal + Semantic			
Predictors	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI
(Intercept)	12.800 ***	-0.055	10.692 – 14.908	-0.138 – 0.029	12.431 ***	-0.057	10.181 – 14.681	-0.140 – 0.026	12.604 ***	-0.056	10.351 – 14.856	-0.139 – 0.027	12.528 ***	-0.058	10.262 – 14.794	-0.142 – 0.026
Wave2	-0.321 *	-0.055	-0.578 – -0.063	-0.099 – -0.011	-0.330 *	-0.057	-0.594 – -0.066	-0.102 – -0.011	-0.328 *	-0.056	-0.592 – -0.064	-0.102 – -0.011	-0.287 *	-0.049	-0.552 – -0.022	-0.095 – -0.004
Wave3	-0.543 ***	-0.093	-0.809 – -0.276	-0.139 – -0.047	-0.556 ***	-0.096	-0.823 – -0.289	-0.141 – -0.050	-0.562 ***	-0.097	-0.829 – -0.295	-0.142 – -0.051	-0.536 ***	-0.092	-0.804 – -0.268	-0.138 – -0.046
age	-0.081 ***	-0.168	-0.109 – -0.053	-0.227 – -0.110	-0.076 ***	-0.158	-0.104 – -0.048	-0.216 – -0.100	-0.079 ***	-0.164	-0.107 – -0.051	-0.222 – -0.106	-0.076 ***	-0.157	-0.104 – -0.047	-0.216 – -0.098
education years	0.002	0.001	-0.068 – 0.071	-0.058 – 0.061	0.004	0.004	-0.065 – 0.073	-0.056 – 0.063	0.004	0.003	-0.065 – 0.073	-0.056 – 0.063	-0.000	-0.000	-0.070 – 0.069	-0.060 – 0.059
gender2	1.336 ***	0.229	0.654 – 2.018	0.112 – 0.347	1.386 ***	0.238	0.707 – 2.066	0.121 – 0.355	1.369 ***	0.235	0.688 – 2.050	0.118 – 0.352	1.342 ***	0.230	0.656 – 2.029	0.113 – 0.348
income	-0.811 ***	-0.206	-1.046 – -0.576	-0.265 – -0.146	-0.811 ***	-0.206	-1.045 – -0.577	-0.265 – -0.146	-0.809 ***	-0.205	-1.044 – -0.574	-0.265 – -0.146	-0.803 ***	-0.204	-1.038 – -0.569	-0.263 – -0.144
tabacco days	0.055 **	0.057	0.014 – 0.097	0.014 – 0.100	0.056 **	0.057	0.014 – 0.097	0.015 – 0.100	0.055 **	0.056	0.013 – 0.096	0.014 – 0.099	0.048 *	0.049	0.006 – 0.089	0.006 – 0.092
urbanization	-0.244	-0.038	-0.615 – 0.126	-0.096 – 0.020	-0.290	-0.045	-0.659 – 0.080	-0.103 – 0.013	-0.267	-0.042	-0.638 – 0.104	-0.100 – 0.016	-0.276	-0.043	-0.647 – 0.095	-0.101 – 0.015
Average daily count of apps					0.003 **	0.066	0.001 – 0.004	0.025 – 0.107								
Average daily count of URLs					0.001	0.014	-0.002 – 0.004	-0.025 – 0.054								
Ratio of active days					-0.315	-0.013	-1.273 – 0.644	-0.052 – 0.026	-0.138	-0.006	-1.132 – 0.856	-0.046 – 0.035	-0.244	-0.010	-1.266 – 0.778	-0.052 – 0.032
Time spent on calls					0.002	0.024	-0.001 – 0.005	-0.014 – 0.063	0.002	0.028	-0.001 – 0.005	-0.010 – 0.067	0.002	0.023	-0.001 – 0.005	-0.015 – 0.061
E: count of URLs									0.000	0.002	-0.001 – 0.001	-0.039 – 0.044	0.000	0.004	-0.001 – 0.001	-0.042 – 0.049
M: total duration									0.000	0.015	-0.000 – 0.001	-0.029 – 0.059				
N: count of apps									0.001 **	0.042	0.000 – 0.001	0.011 – 0.074	0.001 **	0.042	0.000 – 0.001	0.010 – 0.074
N: count of URLs									0.001	0.015	-0.001 – 0.003	-0.021 – 0.051				
N: total duration									-0.000	-0.021	-0.001 – 0.000	-0.062 – 0.020	-0.000	-0.027	-0.001 – 0.000	-0.067 – 0.014
chat and messaging: count of URLs													-0.001	-0.004	-0.007 – 0.006	-0.037 – 0.030
chat and messaging: total duration													0.001 **	0.046	0.000 – 0.002	0.013 – 0.078
email: count of URLs													-0.002	-0.032	-0.005 – 0.001	-0.078 – 0.015
email: total duration													-0.001	-0.012	-0.002 – 0.001	-0.050 – 0.025
gambling: count of apps													0.005	0.003	-0.073 – 0.083	-0.045 – 0.050
gambling: total duration													0.002	0.005	-0.007 – 0.010	-0.025 – 0.036
games: count of apps													0.001	0.016	-0.001 – 0.002	-0.029 – 0.061
games: count of URLs													0.006	0.017	-0.003 – 0.015	-0.008 – 0.043
health: count of apps													0.000	0.004	-0.003 – 0.003	-0.033 – 0.040
health: count of URLs													-0.010	-0.016	-0.030 – 0.009	-0.046 – 0.014
job related: count of apps													0.003	0.032	-0.002 – 0.007	-0.018 – 0.083
job related: count of URLs													0.024 ***	0.075	0.011 – 0.037	0.033 – 0.117
message boards and forums: count of URLs													-0.001	-0.002	-0.016 – 0.014	-0.030 – 0.025
message boards and forums: total duration													-0.001	-0.013	-0.005 – 0.002	-0.050 – 0.023
search engines and portals: total duration													-0.000	-0.001	-0.002 – 0.002	-0.037 – 0.035
shopping: count of apps													0.000	0.004	-0.002 – 0.002	-0.043 – 0.051
social networking: count of URLs													0.000	0.001	-0.002 – 0.002	-0.039 – 0.042
social networking: total duration													-0.000	-0.008	-0.001 – 0.001	-0.052 – 0.037
streaming media: count of URLs													0.001	0.004	-0.008 – 0.010	-0.025 – 0.033
streaming media: total duration													0.001	0.033	-0.000 – 0.002	-0.008 – 0.073
Random Effects																
σ^2	6.64				6.65				6.66				6.64			
τ00	23.67 pid				23.37 pid				23.48 pid				23.28 pid			
ICC	0.78				0.78				0.78				0.78			
N	898 pid				898 pid				898 pid				898 pid			
Observations	2341				2341				2341				2341			
Marginal R^2 / Conditional R^2	0.091 / 0.801				0.098 / 0.800				0.095 / 0.800				0.105 / 0.801			

* p0.05 ** p0.01 *** p0.001

Mobile PHQ-9

Table G2: Desktop: depression score (PHQ-9) hierarchical models results. For each model, the pre-selected features reported in Table 18 are further screened using a VIF threshold (VIF $\leq$ 1.5). The analysis is run using the features and survey responses from the first three waves of the WebWell longitudinal study. The standardized β coefficients (st.Beta) are calculated using the Gelman approach [45].

	Sociodemographics				Sociodemographics + Aggregate Volume				Sociodemographics + Aggregate Volume + Temporal				Sociodemographics + Aggregate Volume + Temporal + Semantic			
Predictors	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI
(Intercept)	15.097 ***	0.041	13.122 – 17.072	-0.025 – 0.107	15.057 ***	0.041	13.055 – 17.059	-0.026 – 0.107	15.100 ***	0.040	13.121 – 17.079	-0.026 – 0.106	15.023 ***	0.039	13.032 – 17.015	-0.027 – 0.106
Wave2	-0.132	-0.024	-0.390 – 0.126	-0.070 – 0.023	-0.130	-0.023	-0.391 – 0.132	-0.070 – 0.024	-0.125	-0.022	-0.385 – 0.136	-0.069 – 0.024	-0.111	-0.020	-0.371 – 0.149	-0.067 – 0.027
Wave3	-0.349 *	-0.063	-0.628 – -0.070	-0.113 – -0.013	-0.343 *	-0.062	-0.623 – -0.063	-0.112 – -0.011	-0.337 *	-0.061	-0.617 – -0.056	-0.111 – -0.010	-0.327 *	-0.059	-0.608 – -0.046	-0.109 – -0.008
age	-0.112 ***	-0.247	-0.138 – -0.086	-0.305 – -0.189	-0.112 ***	-0.247	-0.139 – -0.086	-0.306 – -0.189	-0.112 ***	-0.248	-0.139 – -0.086	-0.306 – -0.190	-0.111 ***	-0.245	-0.138 – -0.085	-0.304 – -0.187
education years	-0.039	-0.035	-0.106 – 0.027	-0.094 – 0.024	-0.040	-0.035	-0.106 – 0.027	-0.094 – 0.024	-0.038	-0.033	-0.104 – 0.029	-0.092 – 0.026	-0.041	-0.036	-0.108 – 0.026	-0.095 – 0.023
gender2	0.072	0.013	-0.191 – 0.336	-0.034 – 0.060	0.074	0.013	-0.189 – 0.338	-0.034 – 0.061	0.076	0.014	-0.188 – 0.340	-0.034 – 0.061	0.075	0.014	-0.189 – 0.340	-0.034 – 0.061
income	-0.694 ***	-0.190	-0.908 – -0.479	-0.249 – -0.131	-0.690 ***	-0.189	-0.905 – -0.475	-0.248 – -0.130	-0.700 ***	-0.192	-0.915 – -0.485	-0.251 – -0.133	-0.703 ***	-0.193	-0.920 – -0.486	-0.252 – -0.133
tabacco days	0.084 ***	0.089	0.040 – 0.128	0.043 – 0.135	0.084 ***	0.089	0.040 – 0.128	0.043 – 0.135	0.086 ***	0.091	0.042 – 0.130	0.044 – 0.137	0.087 ***	0.092	0.042 – 0.131	0.045 – 0.139
urbanization	-0.369 *	-0.060	-0.723 – -0.015	-0.118 – -0.002	-0.372 *	-0.061	-0.727 – -0.018	-0.119 – -0.003	-0.368 *	-0.060	-0.722 – -0.014	-0.118 – -0.002	-0.363 *	-0.059	-0.717 – -0.008	-0.117 – -0.001
Average daily count of URLs					0.000	0.015	-0.001 – 0.001	-0.029 – 0.058								
Ratio of active days					-0.048	-0.003	-0.754 – 0.658	-0.042 – 0.037								
A: count of URLs									0.000	0.018	-0.000 – 0.000	-0.027 – 0.063				
M: total duration									-0.000	-0.010	-0.001 – 0.000	-0.054 – 0.033				
N: total duration									-0.000	-0.023	-0.001 – 0.000	-0.070 – 0.023				
E: count of URLs													0.000	0.020	-0.000 – 0.000	-0.031 – 0.070
M: count of URLs													0.000	0.008	-0.000 – 0.000	-0.034 – 0.050
N: count of URLs													-0.000	-0.037	-0.001 – 0.000	-0.081 – 0.008
chat and messaging: total duration													0.001	0.011	-0.002 – 0.003	-0.034 – 0.056
email: total duration													-0.001	-0.016	-0.003 – 0.001	-0.062 – 0.030
gambling: count of URLs													-0.000	-0.007	-0.003 – 0.002	-0.053 – 0.039
games: total duration													-0.000	-0.009	-0.002 – 0.001	-0.052 – 0.033
health: count of URLs													-0.003	-0.016	-0.007 – 0.002	-0.046 – 0.013
job related: total duration													0.012	0.006	-0.040 – 0.065	-0.019 – 0.031
message boards and forums: total duration													0.014 **	0.051	0.005 – 0.024	0.018 – 0.084
search engines and portals: total duration													0.000	0.004	-0.001 – 0.001	-0.054 – 0.063
shopping: total duration													0.001	0.026	-0.000 – 0.003	-0.007 – 0.058
social networking: total duration													-0.000	-0.016	-0.001 – 0.000	-0.062 – 0.030
streaming media: total duration													0.000	0.003	-0.001 – 0.001	-0.037 – 0.044
Random Effects																
σ^2	6.51				6.51				6.52				6.51			
τ00	22.09 pid				22.15 pid				22.07 pid				22.08 pid			
ICC	0.77				0.77				0.77				0.77			
N	946 pid				946 pid				946 pid				946 pid			
Observations	2283				2283				2283				2283			
Marginal R^2 / Conditional R^2	0.096 / 0.794				0.096 / 0.795				0.096 / 0.794				0.100 / 0.795			

* p0.05 ** p0.01 *** p0.001

Desktop PHQ-9

Table G3: Mobile: suicide risk score (PHQ-9-Q9) hierarchical models results. For each model, the pre-selected features reported in Table 18 are further screened using a VIF threshold (VIF ≤ 1.5). The analysis is run using the features and survey responses from the first three waves of the WebWell longitudinal study. The standardized β coefficients (st.Beta) are calculated using the Gelman approach [45].

	Sociodemographics				Sociodemographics + Aggregate Volume				Sociodemographics + Aggregate Volume + Temporal				Sociodemographics + Aggregate Volume + Temporal + Semantic			
Predictors	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI
(Intercept)	0.9889 ***	0.0545	0.7472 – 1.2305	-0.0298 – 0.1387	0.9747 ***	0.0529	0.7102 – 1.2392	-0.0312 – 0.1370	0.9937 ***	0.0555	0.7294 – 1.2580	-0.0286 – 0.1397	0.9853 ***	0.0486	0.7204 – 1.2503	-0.0354 – 0.1326
Wave2	-0.0237	-0.0349	-0.0636 – 0.0162	-0.0935 – 0.0238	-0.0242	-0.0356	-0.0648 – 0.0164	-0.0952 – 0.0240	-0.0241	-0.0354	-0.0647 – 0.0165	-0.0951 – 0.0242	-0.0169	-0.0249	-0.0575 – 0.0236	-0.0845 – 0.0347
Wave3	-0.0286	-0.0420	-0.0698 – 0.0127	-0.1026 – 0.0187	-0.0294	-0.0432	-0.0707 – 0.0119	-0.1039 – 0.0175	-0.0300	-0.0441	-0.0713 – 0.0114	-0.1048 – 0.0167	-0.0275	-0.0405	-0.0688 – 0.0137	-0.1011 – 0.0202
age	-0.0072 ***	-0.1269	-0.0104 – -0.0039	-0.1839 – -0.0699	-0.0069 ***	-0.1216	-0.0101 – -0.0036	-0.1789 – -0.0643	-0.0071 ***	-0.1251	-0.0103 – -0.0038	-0.1820 – -0.0682	-0.0070 ***	-0.1240	-0.0103 – -0.0037	-0.1820 – -0.0660
education years	-0.0049	-0.0361	-0.0128 – 0.0030	-0.0943 – 0.0222	-0.0049	-0.0360	-0.0128 – 0.0030	-0.0942 – 0.0222	-0.0051	-0.0377	-0.0130 – 0.0028	-0.0958 – 0.0205	-0.0050	-0.0368	-0.0129 – 0.0029	-0.0948 – 0.0212
gender2	-0.0520	-0.0764	-0.1300 – 0.0260	-0.1910 – 0.0381	-0.0489	-0.0718	-0.1268 – 0.0290	-0.1863 – 0.0427	-0.0524	-0.0770	-0.1302 – 0.0255	-0.1914 – 0.0375	-0.0476	-0.0699	-0.1259 – 0.0308	-0.1851 – 0.0453
income	-0.0612 ***	-0.1329	-0.0881 – -0.0344	-0.1912 – -0.0746	-0.0613 ***	-0.1330	-0.0881 – -0.0345	-0.1912 – -0.0748	-0.0623 ***	-0.1353	-0.0891 – -0.0355	-0.1935 – -0.0770	-0.0630 ***	-0.1368	-0.0897 – -0.0363	-0.1947 – -0.0789
tabacco days	0.0029	0.0255	-0.0025 – 0.0084	-0.0224 – 0.0735	0.0028	0.0243	-0.0027 – 0.0082	-0.0236 – 0.0722	0.0029	0.0254	-0.0026 – 0.0083	-0.0225 – 0.0734	0.0031	0.0276	-0.0023 – 0.0086	-0.0204 – 0.0757
urbanization	-0.0153	-0.0205	-0.0577 – 0.0271	-0.0771 – 0.0362	-0.0181	-0.0242	-0.0605 – 0.0243	-0.0809 – 0.0325	-0.0176	-0.0236	-0.0601 – 0.0248	-0.0803 – 0.0332	-0.0190	-0.0254	-0.0613 – 0.0232	-0.0819 – 0.0310
Average daily count of apps					0.0001	0.0251	-0.0001 – 0.0003	-0.0225 – 0.0726								
Average daily count of URLs					0.0002	0.0164	-0.0003 – 0.0006	-0.0296 – 0.0623								
Ratio of active days					-0.0350	-0.0123	-0.1644 – 0.0944	-0.0576 – 0.0330	-0.0325	-0.0114	-0.1665 – 0.1016	-0.0583 – 0.0356	-0.0185	-0.0065	-0.1553 – 0.1184	-0.0544 – 0.0414
Time spent on calls					0.0004	0.0445	-0.0000 – 0.0009	-0.0002 – 0.0892	0.0004 *	0.0448	0.0000 – 0.0009	0.0003 – 0.0894	0.0004	0.0438	-0.0000 – 0.0009	-0.0007 – 0.0884
E: count of URLs									0.0000	0.0164	-0.0001 – 0.0001	-0.0320 – 0.0648	0.0000	0.0182	-0.0001 – 0.0001	-0.0350 – 0.0715
M: total duration									0.0000	0.0258	-0.0000 – 0.0001	-0.0253 – 0.0770				
N: count of apps									0.0000	0.0057	-0.0001 – 0.0001	-0.0330 – 0.0445	0.0000	0.0027	-0.0001 – 0.0001	-0.0360 – 0.0414
N: count of URLs									0.0000	0.0071	-0.0002 – 0.0003	-0.0372 – 0.0514				
N: total duration									-0.0000	-0.0282	-0.0001 – 0.0000	-0.0765 – 0.0201	-0.0000	-0.0177	-0.0001 – 0.0000	-0.0655 – 0.0301
chat and messaging: count of URLs													0.0008	0.0346	-0.0001 – 0.0017	-0.0062 – 0.0755
chat and messaging: total duration													0.0002 **	0.0618	0.0001 – 0.0003	0.0217 – 0.1019
email: count of URLs													0.0000	0.0050	-0.0004 – 0.0005	-0.0480 – 0.0580
email: total duration													0.0000	0.0032	-0.0002 – 0.0003	-0.0414 – 0.0479
gambling: count of apps													-0.0015	-0.0078	-0.0114 – 0.0084	-0.0593 – 0.0437
gambling: total duration													-0.0003	-0.0096	-0.0016 – 0.0010	-0.0473 – 0.0282
games: count of apps													-0.0002	-0.0465	-0.0004 – 0.0000	-0.0966 – 0.0035
games: count of URLs													0.0006	0.0152	-0.0007 – 0.0020	-0.0174 – 0.0478
health: count of apps													0.0004 *	0.0468	0.0000 – 0.0008	0.0034 – 0.0901
health: count of URLs													-0.0014	-0.0180	-0.0043 – 0.0015	-0.0554 – 0.0195
job related: count of apps													0.0004	0.0415	-0.0001 – 0.0010	-0.0116 – 0.0946
job related: count of URLs													0.0019 *	0.0499	0.0001 – 0.0036	0.0023 – 0.0975
message boards and forums: count of URLs													-0.0012	-0.0188	-0.0034 – 0.0010	-0.0535 – 0.0158
message boards and forums: total duration													-0.0002	-0.0124	-0.0007 – 0.0004	-0.0553 – 0.0305
search engines and portals: total duration													-0.0002	-0.0351	-0.0004 – 0.0000	-0.0784 – 0.0082
shopping: count of apps													0.0001	0.0159	-0.0002 – 0.0003	-0.0360 – 0.0678
social networking: count of URLs													-0.0002	-0.0221	-0.0005 – 0.0002	-0.0704 – 0.0261
social networking: total duration													-0.0000	-0.0158	-0.0001 – 0.0001	-0.0659 – 0.0344
streaming media: count of URLs													-0.0000	-0.0003	-0.0013 – 0.0013	-0.0360 – 0.0353
streaming media: total duration													0.0000	0.0096	-0.0001 – 0.0002	-0.0369 – 0.0561
Random Effects																
σ^2	0.16				0.16				0.16				0.16			
τ00	0.28 pid				0.28 pid				0.28 pid				0.27 pid			
ICC	0.63				0.63				0.63				0.63			
N	898 pid				898 pid				898 pid				898 pid			
Observations	2341				2341				2341				2341			
Marginal R^2 / Conditional R^2	0.034 / 0.646				0.038 / 0.646				0.038 / 0.645				0.052 / 0.647			

* p0.05 ** p0.01 *** p0.001

Mobile PHQ-9-Q9

Table G4: Desktop: suicide risk score (PHQ-9-Q9) hierarchical models results. For each model, the pre-selected features reported in Table 18 are further screened using a VIF threshold (VIF $\leq$ 1.5). The analysis is run using the features and survey responses from the first three waves of the WebWell longitudinal study. The standardized β coefficients (st.Beta) are calculated using the Gelman approach [45].

	Sociodemographics				Sociodemographics + Aggregate Volume				Sociodemographics + Aggregate Volume + Temporal				Sociodemographics + Aggregate Volume + Temporal + Semantic			
Predictors	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI	Beta	std. Beta	CI	std. CI
(Intercept)	1.001 ***	0.005	0.774 – 1.229	-0.065 – 0.075	0.993 ***	0.004	0.761 – 1.224	-0.066 – 0.074	1.001 ***	0.003	0.773 – 1.228	-0.067 – 0.072	1.020 ***	0.003	0.790 – 1.250	-0.067 – 0.073
Wave2	-0.006	-0.008	-0.044 – 0.033	-0.067 – 0.051	-0.005	-0.007	-0.044 – 0.034	-0.067 – 0.052	-0.004	-0.006	-0.042 – 0.035	-0.065 – 0.054	-0.004	-0.007	-0.043 – 0.034	-0.066 – 0.053
Wave3	-0.026	-0.040	-0.068 – 0.015	-0.104 – 0.024	-0.025	-0.039	-0.067 – 0.016	-0.102 – 0.025	-0.023	-0.036	-0.065 – 0.018	-0.100 – 0.028	-0.025	-0.038	-0.067 – 0.017	-0.102 – 0.026
age	-0.009 ***	-0.163	-0.012 – -0.006	-0.219 – -0.106	-0.009 ***	-0.164	-0.012 – -0.006	-0.221 – -0.106	-0.009 ***	-0.164	-0.012 – -0.006	-0.221 – -0.107	-0.009 ***	-0.167	-0.012 – -0.006	-0.226 – -0.109
education years	-0.007	-0.052	-0.015 – 0.001	-0.110 – 0.005	-0.007	-0.052	-0.015 – 0.001	-0.110 – 0.005	-0.007	-0.049	-0.014 – 0.001	-0.107 – 0.009	-0.007	-0.052	-0.015 – 0.001	-0.110 – 0.006
gender2	0.013	0.019	-0.026 – 0.051	-0.040 – 0.078	0.013	0.020	-0.026 – 0.051	-0.039 – 0.079	0.013	0.021	-0.025 – 0.052	-0.038 – 0.080	0.013	0.020	-0.026 – 0.052	-0.040 – 0.079
income	-0.054 ***	-0.126	-0.079 – -0.029	-0.184 – -0.068	-0.054 ***	-0.125	-0.078 – -0.029	-0.183 – -0.067	-0.055 ***	-0.129	-0.080 – -0.031	-0.187 – -0.072	-0.055 ***	-0.127	-0.080 – -0.030	-0.186 – -0.069
tabacco days	0.007 *	0.059	0.001 – 0.012	0.009 – 0.109	0.007 *	0.059	0.001 – 0.012	0.009 – 0.109	0.007 *	0.064	0.002 – 0.013	0.014 – 0.114	0.007 *	0.063	0.001 – 0.013	0.012 – 0.113
urbanization	-0.012	-0.016	-0.052 – 0.029	-0.073 – 0.040	-0.012	-0.017	-0.053 – 0.029	-0.073 – 0.040	-0.011	-0.016	-0.052 – 0.029	-0.072 – 0.041	-0.014	-0.019	-0.055 – 0.027	-0.076 – 0.038
Average daily count of URLs					0.000	0.014	-0.000 – 0.000	-0.035 – 0.063								
Ratio of active days					0.005	0.002	-0.091 – 0.100	-0.043 – 0.048								
A: count of URLs									0.000	0.042	-0.000 – 0.000	-0.009 – 0.092				
M: total duration									-0.000	-0.028	-0.000 – 0.000	-0.077 – 0.021				
N: total duration									-0.000	-0.044	-0.000 – 0.000	-0.095 – 0.007				
E: count of URLs													0.000	0.012	-0.000 – 0.000	-0.044 – 0.068
M: count of URLs													0.000	0.011	-0.000 – 0.000	-0.038 – 0.059
N: count of URLs													-0.000	-0.001	-0.000 – 0.000	-0.051 – 0.049
chat and messaging: total duration													0.000	0.016	-0.000 – 0.000	-0.034 – 0.065
email: total duration													0.000	0.000	-0.000 – 0.000	-0.051 – 0.052
gambling: count of URLs													-0.000	-0.005	-0.000 – 0.000	-0.057 – 0.047
games: total duration													-0.000	-0.020	-0.000 – 0.000	-0.069 – 0.029
health: count of URLs													-0.001	-0.030	-0.001 – 0.000	-0.066 – 0.006
job related: total duration													-0.001	-0.006	-0.009 – 0.006	-0.037 – 0.025
message boards and forums: total duration													0.000	0.010	-0.001 – 0.002	-0.029 – 0.049
search engines and portals: total duration													-0.000	-0.037	-0.000 – 0.000	-0.096 – 0.021
shopping: total duration													0.000	0.008	-0.000 – 0.000	-0.031 – 0.047
social networking: total duration													0.000	0.014	-0.000 – 0.000	-0.036 – 0.064
streaming media: total duration													-0.000	-0.018	-0.000 – 0.000	-0.065 – 0.028
Random Effects																
σ^2	0.15				0.15				0.15				0.15			
τ00	0.26 pid				0.27 pid				0.26 pid				0.27 pid			
ICC	0.64				0.64				0.64				0.64			
N	946 pid				946 pid				946 pid				946 pid			
Observations	2283				2283				2283				2283			
Marginal R^2 / Conditional R^2	0.044 / 0.660				0.044 / 0.660				0.047 / 0.660				0.047 / 0.661			

* p0.05 ** p0.01 *** p0.001

Desktop PHQ-9-Q9