Individual-level Anxiety Detection and Prediction from Longitudinal YouTube and Google Search Engagement Logs

Anis Zaman[1], **Boyu Zhang** [1], **Vincent Silenzio** [2], **Henry Kautz** [1], **Ehsan Hoque** [1]
[1] Department of Computer Science, University of Rochester, NY, USA
[2] Department of Urban-Global Public Health, Rutgers University, NJ, USA
[azaman2, kautz, mehoque]@cs.rochester.edu, bzhang25@u.rochester.edu, vincent.silenzio@rutgers.edu,

Abstract

Anxiety disorder is one of the most prevalent mental health conditions globally, arising from complex interactions of biological and environmental factors and severely interfering with one's ability to lead normal life activities. Current methods for detecting anxiety heavily rely on in-person interviews, which can be expensive, time-consuming, and blocked by social stigmas. We propose an alternative method to identify individuals with anxiety and further estimate their levels of anxiety using personal online activity histories from YouTube and the Google Search engine. We ran a longitudinal study and collected multiple rounds of anonymized YouTube and Google Search logs from volunteering participants, along with their clinically validated ground-truth anxiety assessment scores. We then developed explainable features that capture both the temporal and contextual aspects of online behaviors. Using those, we were able to train models that can (i) identify individuals having anxiety disorder (ii) assess the level of anxiety by accurately predicting the gold-standard Generalized Anxiety Disorder 7-item scores based on the ubiquitous individual-level online engagement data. Our proposed anxiety assessment framework can be deployed in clinical settings, empowering care providers to learn about anxiety disorders of patients non-invasively.

Introduction

According to the World Health Organization (WHO), 1 in 13 people suffers from anxiety globally, making it one of the most prevalent mental health concerns. In the United States, it is the second leading cause of disability among all psychiatric disorders (Whiteford et al. 2013). Nearly 40 million people (age 18 and older) experience anxiety disorder in any given year, yet only 35.9% of those suffered received treatments[1]. A study in 2017 reported that the level of anxiety among young adolescents has been gradually increasing in recent years (Calling et al. 2017).

The population most vulnerable to anxiety disorder is the students in high school and early college years. A report by the American College Health Association in 2018 stated that 63% of college students in the U.S. felt overwhelming anxiety during the last 12 months, and only 23% of these students were either diagnosed or treated for an anxiety disorder by a professional mental healthcare provider (Association. 2018). During the early days of college, students are separated from their traditional support system and find themselves in challenging social and academic settings such as living with roommates, developing independent identities, making new friends, managing heavy workloads, etc. All these experiences induce spikes in anxiety from time to time (Purdon et al. 2001) Furthermore, it has been reported that anxiety disorders are significantly associated with other medical and psychiatric comorbidities (Costello, Egger, and Angold 2005). Despite such a high prevalence of anxiety among young adolescents, current methods for detecting anxiety disorders consist of self-assessment surveys and in-person interviews, which can be time-consuming, expensive, lack of precision, and hampered by factors such as fear, concealing information, and social stigma related to the mental health issue.

User engagements with online platforms are major components in the lives of young adults (Kaplan and Haenlein 2010). On average, an internet user spent the equivalent of more than 100 days online during the last 12 months (The Next Web 2020). It has been reported that 81% of U.S internet users aging between 15 to 25 use YouTube[2] regularly. Besides, an average internet user uses Google Search at least once a day, and many search dozens of times a day[3]. Extensive studies have been done trying to correlate mental health issues with popular public social media data such as Facebook (Ophir, Asterhan, and Schwarz 2020) and Twitter (Coppersmith, Dredze, and Harman 2014; Gopalakrishna Pillai, Thelwall, and Orasan 2018), yet they may fail to cover people who interact infrequently with social media or post false positive impressions publicly (Gil-Or1, Levi-Belzm, and Turel 2015). In contrast, individual-level search and YouTube logs are ubiquitous and private for each user and are less likely to be subject to self-censorship. A group of researchers has shown that search logs can be used as a proxy for detecting mental health issues (Adler et al. 2019; Jimenez, Santed-Germán, and Ramos 2020; Zaman et al. 2019). We draw inspirations from these prior works

[1] https://adaa.org/understanding-anxiety

[2] https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[3] https://bit.ly/382vgWD

and hypothesize that private Google Search engine logs and YouTube histories can leave a detailed digital trace of the mental health states of users and be used as a proxy to assess the level of anxiety for individuals.

In this work, we propose a framework that leverages individual-level online activity logs, in particular, Google Search and YouTube activity histories, to *identify* individuals with anxiety disorder and further *predict* their level of anxiety. We ran a longitudinal study to gather two rounds of data, with 5 months in-between, from a college cohort. During each round, participants shared their anonymized online activity histories along with their answers to a clinically validated questionnaire for measuring Generalized Anxiety Disorder (GAD-7) (Spitzer et al. 2006). We then developed an explainable low-dimensional vector representation that captures different aspects of one's online behaviors, including temporal activity patterns based on temporal point processes, time and semantic diversities, and periods of inactivity. Using these feature representations, we trained models that can accurately detect and predict one's level of anxiety from online activities. Unlike (Zaman et al. 2019) who merely focused on mental health issue *detection* such as self-esteem from Google Search histories, our data incorporates both Google Search as well as YouTube activities history, and our novel two rounds of data facilitate both the *detection* and *prediction* tasks. Furthermore, we conduct our experiment with a framework that fits possible real-world applications. We envision our work as an important step towards empowering caregivers to better understand and engage with their patients non-invasively through passive data and ubiquitous computing.

Related Work

Public social media, blogs, and forums have become popular data sources for researchers to study the prevalence of mental health conditions. (Seabrook, Kern, and Rickard 2016) showed that the usage of social media sites correlates with user depression and anxiety. Twitter has been used to detect insomnia (Jamison-Powell et al. 2012), suicidal ideations (De Choudhury et al. 2016), depressed individuals (De Choudhury et al. 2013), and languages related to depression and PTSD (Reece et al. 2017; Coppersmith et al. 2015).Besides, Facebook status can be used to predict postpartum depression (De Choudhury et al. 2014) and monitor depression (Schwartz et al. 2014). Other researchers leveraged Reddit to study mental distress among adolescents (Bagroy, Kumaraguru, and De Choudhury 2017). De Choudhury et al. provides a comprehensive overview of the role of social media in mental health researches (De Choudhury, Counts, and Horvitz 2013) and evaluation methodologies (Chancellor and De Choudhury 2020). However, Social media users constitutes only a fraction of the general population. Only a small number of them, with particular personalities or demographics, acts out on public platforms, revealing signs of mental health struggles. Hence, findings based on social media platforms may not generalize to the majority of the population.

One data source that can capture in-the-moment thoughts and feelings of a broad range of people is search engine log,

Figure 1: Collecting online data from an individual.

which may fill in the gap for continuous monitoring applications (Mohr, Zhang, and Schueller 2017). Researchers have used population-level search engine logs from Google Trends to monitor depression and suicide-related behaviors (McCarthy 2010; Sueki 2011; Yang et al. 2011; Gunn III and Lester 2013), identify seasonality in seeking mental health information (Ayers et al. 2013), and show heavy usages for screening diseases (Paparrizos, White, and Horvitz 2016a) such as pancreatic cancer (Paparrizos, White, and Horvitz 2016b). A comprehensive review of the usage of Google Trends in the healthcare domain has been provided by (Nuti et al. 2014). A crucial difference between these previous works and ours is that we aim to accurately predict the mental health of particular individuals, not general populations. Unlike population-level online engagement logs in Google Trends, our individual-level activity logs are more likely to fit the fabric of one's daily life experience.

Data

The longitudinal data collected for this work consisted of individual-level Google Search logs, YouTube history, and clinical survey responses that are very personal and sensitive in nature. Similar to (Zaman et al. 2019), we leveraged a cloud-based data collection process using Google Takeout[4], a web interface that enables Google product users to export their Google Search and YouTube activity histories. Our cloud-based data collection pipeline (see Figure 1) is HIPAA-compliant and has been thoroughly vetted by the Institutional Review Board (IRB) of our institution in order to ensure the privacy and safety of subjects.

Study Recruitment Procedure

The study ran for 5 months starting in August, 2019. Participation was voluntary, and one needed to be at least 18-year-old and have a Google account to qualify for the study. The recruitment procedure was designed as an one-on-one interview. During the recruitment, participants answered the 7-item Generalized Anxiety Disorder questionnaire, a clinically validated tool for assessing anxiety disorder, in addition to their GPA, gender, and demographics. Following that, participants signed in to Google Takeout with their Google accounts and initiated the Google Search and YouTube activity history data download process. Before the data was shared with the research team, all sensitive information such as name, email, phone number, social security, and financial information (banking and credit card) was redacted and anonymized using Google's Data Loss Prevention (DLP) API (Kiang and Bailon 2016; Kim and Paek 2016).

In total, we collected two rounds of data. The recruitment procedure above was performed during each round. In

[4]http://takeout.google.com/

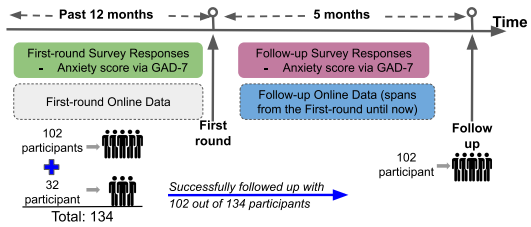


Figure 2: Two rounds of data collection.

Individuals	First round	Follow up
# Anxious	62	44
# Not anxious	72	58
Total	134	102

(a) Distributions of participants with and without anxiety conditions during the first and follow-up rounds.

# participants, N = 134				
White	53 (39.6%)	Mean Age	20.99	
African Americans	48 (35.8%)	Male	57 (42.5%)	
Other	33 (24.6%)	Female	77 (57.5%)	

(b) Demographics of the participants.

Figure 3: Study population breakdown

August 2019, 134 qualified college college students participated in the first round. For the rest of the paper, we will refer this round of data as the *first-round* data.

Five months later, we invited all 134 participants from *first-round* for follow-up and were able to follow up with 102 individuals. We collected their Google and YouTube activity histories again, along with the survey responses for the second time. For the rest of the paper, we will refer to data collected in the second round as the *follow-up* data. Therefore, there are in total 102 people participated in both rounds and $134 - 102 = 32$ people participated only in the *first-round*. The overall recruitment timeline and participant statistics are shown in Figure 2. All participants were compensated with 10$ Amazon gift cards during each round of participation. 42.5% of our participants are male, and 57.5% are female. No participant reported non-binary gender though we gave the options. Figure 3(b) presents a comprehensive breakdown of the demographics of the study population.

Ground Truth via Survey

The ground truth about one's anxiety disorder was measured using the Generalized Anxiety Disorder (GAD-7) (Spitzer et al. 2006), a clinically validated questionnaire (7 questions[5]) which has been reported to be accurate in accessing the severity of anxiety (Swinson 2006). The questions in GAD-7 were prefixed with a text for the temporal context. For example, *Over the last five months, how often have you been bothered by the following problems?* The responses were converted to an anxiety score on a 21-point scale. GAD-7 is a commonly used in clinical diagnosis where scores of 5, 10, and 15 are treated as cutoffs for mild, moderate, and severe anxiety levels, respectively. Further follow-up and evaluation are recommended for someone with an anxiety score > 9 (Williams 2014). We used the recommended score of 9 as a cutoff to label individuals with anxiety disorder. In this work, any individual with a GAD-7 score > 9 is labelled as *Anxious*, and someone with a score ≤ 9 is labelled as *Not-anxious*. Figure 3(a) shows the breakdown after the anxiety cutoff. Figure 4 shows the distribution and changes of anxiety scores for all the participants who participated in both the *first-round* and the *follow-up*. We observed that the anxiety score increased for 35 individuals, decreased for 46 individuals, and remain unchanged for 21 individuals. 13 participants had a change in GAD-7 score that is clinically significant (the absolute value of the change ≥ 5) during the 5 months of study.

[5]https://www.mdcalc.com/gad-7-general-anxiety-disorder-7

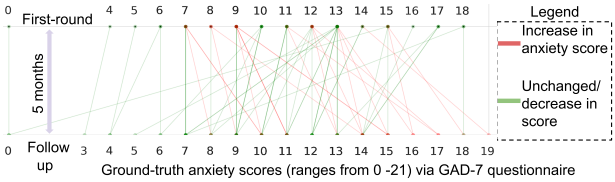


Figure 4: GAD-7 during the *first-round* and *follow-up*.

YouTube & Google Search History

We collect individual-level YouTube and Google Search engine logs via the Google Takeout platform. As long as one is logged into their Google account, regardless of which device is used, Google ties all online activities into a single repository which is accessible through the Takeout interface. For every person, the online activity history spanned (on average) over 5.7 years. In total, 1,966,400 Google searches and 1,055,847 YouTube interactions were made by all the participants.

Every engagement on YouTube and Google Search engine is timestamped along with the information whether it is the result of watching or searching. For YouTube activity logs, we use the YouTube API to extract meta-data about the videos that has been watched, which includes the title, category (for context), video length, rating, etc. For Google Search activities, we label every search query text using the content classification feature of the Google Cloud NLP API [6]. Given a query, the API returns one or more possible category labels for the text along with a confidence score. When applicable, we select the category label with the highest confidence. The API returns a hierarchical label for every query, and we consider the root level as the category label for the query. The comprehensive lists of all the categories for both search queries and YouTube videos are listed in (Google 2020) and (TechPostPlus 2019).

Feature Extraction from Online Data

In this section we present the explainable features that we engineered from individual-level engagement logs from YouTube and Google Search engine, an unique data source to capture what may be going through one's mind at any given time. Since online activities are timestamped, one can investigate the weekday/weekend activities, calculate the semantic and temporal properties of these activities, and estimate daily sleeping/resting duration, etc. For example, Figure 5 demonstrates the distribution of activities on YouTube

[6]https://cloud.google.com/natural-language/docs/classifying-text

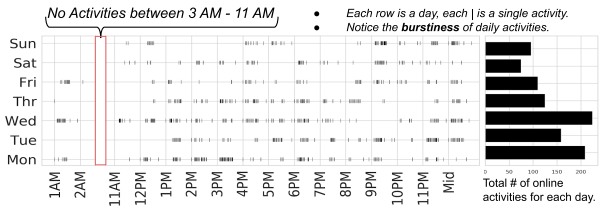


Figure 5: Example online activities distribution from a participant over a week, including both Google Search and YouTube activities.

and Google Search engine over a week for a specific individual in our dataset. Notice that each of the following feature is a scalar and is calculated for each individual participant. In total, we explored four types of features, and each has a number of variants.

Category (C_H) & Time (T_H) Entropy

Every online activity has two components associated with it, namely its category and the timestamp of its occurrence. Drawing inspiration from information theory (Shenkin, Erman, and Mastrandrea 1991), we define category entropy, C_H, as a measure of how diverse an individual's online activities are in terms of the semantic context. For an individual p, based on his/her/their online data, we compute the category entropy in the following way:

$$H_p(Category) = -\sum_{i=1}^{m} P_i \times \log(P_i) \tag{1}$$

where m is the number of distinct categories in the online activities of p, and P_i is the percentage of activities that belong to category i. A high entropy indicates that p interacts more uniformly across different categories online, whereas lower entropy indicates larger inequality in the number of online activities across the categories. Considering that individuals may have different habits during weekdays and weekends, we also calculated the category entropy for weekdays and weekends separately. We include the total, weekday, and weekend category entropy as features for each individual. We denote them as $C_H^{weekday}$, $C_H^{weekend}$, and C_H^{total}.

Similarly, we define time entropy, T_H, as a measure of how diverse an individual's online activities are in terms of when it happens. We define the discrete bins for time entropy as the 24 hours of a day. For a person p, time entropy is computed the same way as C_H above. The difference is that the summation is taken over the 24 hour marks, and P_i is the percentage of activities that happen during hour i. A high entropy indicates that p interacts with YouTube and Google Search engine more uniformly across different times of a day, whereas lower entropy indicates larger inequalities of numbers of online activities between different hours in a day. Also, we obtain the time entropy for weekdays and weekends separately. We denote them as $T_H^{weekday}$, $T_H^{weekend}$, and T_H^{total}.

Online Activities Temporality $\{\gamma, \alpha, \beta\}$

We observed that there is a bursty nature of online activities when plotted on the time axis (see Figure 5) which resulted in clusters of online activities regardless of Google Searches or YouTube histories. In other words, we can view the incidences of online activities as a *Temporal Point Process* and investigate individual-level online behaviors from a temporal point of view, such as the *Inter-event Times* (IETs). We enrich our temporal feature by assuming dependencies between past activities and the next activity. The intuition is that every occurrence of an online activity increases the probability of future online activities, and the probability of the next activity decays with time. Hence, such process, called a self-exciting point process, can be modeled by the Hawkes Process (Hawkes 1971), which has been widely used for modeling online data and social media activities at a population level (Rizoiu et al. 2017). Specifically, we define a univariate Hawkes Process with an exponential decay kernel as

$$\lambda(t) = \gamma + \sum_{t_i<t} \alpha\beta \exp\left(-\beta(t-t_i)\right) \tag{2}$$

where $\lambda(t)$ represents the probability (intensity) of an activity occurs at time t, γ is the background intensity of an activity happens exogenously, α represents the *infectivity factor* which controls the average number of new activities triggered by any past activity, and β is the *decay rate* where $\frac{1}{\beta}$ represents how much time has passed by, on average, between the previous event and the next event. By fitting the above Hawkes Process to each individual online history log, we obtain a unique set of $\{\gamma, \alpha, \beta\}$ for each participant as features. We keep the notations as $\{\gamma, \alpha, \beta\}$ for this set of features.

Inactivity Period $\mathcal{I}$

It has been reported that YouTube is becoming the modern day classroom for students (Fleck et al. 2014) and provides new ways to consume contents for virtually every age groups (Cayari 2011). However, spending too much time on any platform can lead to internet addiction (Hall and Parsons 2001), in particular the YouTube addiction (Moghavvemi et al. 2017) and the compulsive usage of YouTube (Klobas et al. 2018), which are quite prevalent among college population. These previous findings have inspired us consider feature that can be treated as a proxy to capture the time away from internet of each participant, and we call it the inactivity period $\mathcal{I}$.

We focus on periods of time when no Google Search nor YouTube activity was performed of each individual. Given the online activity log of a participant and a duration threshold of k hours, we pick out all the inactive periods longer than k hours and investigate when they happened most frequently. Specifically, for all inactivity periods longer than k hours, we first get the midpoint timestamp for each of them. For example, for an 8-hour inactivity period starting at 11 P.M. and ending at 7 A.M., the midpoint is 3 A.M.

The average inactivity duration of all participants is 9.27 ± 1.17 hours. Hence, we consider $k \in \{8, 9, 10\}$. We found all the midpoint modes fall in-between 5 to 8 A.M., which are most likely to be the middle of sleeping periods. For the inactivity defined here, we are focusing on *when* it occurs *most frequently* for each individual. Hence, it is most

suitable to take the mode of inactivity midpoints. We denote, for threshold $k \in \{8, 9, 10\}$, the inactivity mode features as $\mathcal{I}_8$, $\mathcal{I}_9$, and $\mathcal{I}_{10}$.

Overall, we developed 12 features (including variants) form the online activities of each individual: 3 from each of the Category Entropy C_H, Time Entropy T_H, Online Activities Temporality $\{\gamma, \alpha, \beta\}$, and Inactivity Periods $\mathcal{I}$.

Modeling Anxiety

Following the clinical anxiety score cutoff threshold (Spitzer et al. 2006), participants with GAD-7 score > 9 were labelled as anxious subjects, and those with score ≤ 9 were labelled as non-anxious subjects. Overall, there were 62 out of 134 subjects with anxiety conditions in the *first-round* and 44 out of 102 participants with anxiety conditions during the *follow-up*. Given one's YouTube and Google Search activity history, we explore: (i) Can we identify individuals with anxiety condition through his/her/their online data? (ii) Can we predict anxiety score based on online activities and past anxiety levels?

Notations and Definitions

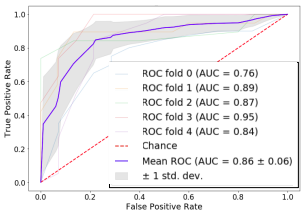
The feature vectors for the *first-round* are extracted using the most recent 12 months of data (the grey box in Figure 2) before the completion of the *first-round* survey. We denote this by $\boldsymbol{x}_1 \in \mathbb{R}^{12}$. Unless mentioned specifically, $\boldsymbol{x}_1$ is the concatenation of all 12 scalar features in the same order for each individual. The corresponding GAD-7 scores, gathered via the survey (the green box in Figure 2) during the *first-round*, are denoted as y_1. Similarly, for the *follow-up* round, the feature vectors are extracted solely from the 5 months of online history data (the blue box in Figure 2) **in-between** the *first-round* and the *follow-up*, and we denote it as $\boldsymbol{x}_2 \in \mathbb{R}^{12}$. The corresponding GAD-7 scores, provided in the *follow-up* survey, (the magenta box in Figure 2), are denoted as y_2. Therefore, there are in total 134 $(\boldsymbol{x}_1, y_1)$ pairs from *first-round* and 102 $(\boldsymbol{x}_2, y_2)$ pairs from *follow-up* (see Figure 2).

Classifying Individuals with Anxiety

Here, we treat the problem as a binary classification task: **given the online activity history, we aim to identify if the participant has anxiety condition**. There are $134 + 102 = 236$ segments ($\boldsymbol{x}_1$ and $\boldsymbol{x}_2$) of online history data in total, regardless of collected in which round or from whom. The respective anxiety scores of individual participants in each round are labels. Formally, we are interested in $P(y \mid \boldsymbol{x})$, where y is the binary anxiety label from the GAD-7 scores cutoff of 9.

We trained random forest (RF) classifiers on this task and performed stratified 5-fold cross-validations. The RF reached an average F1 score of 0.83 ± 0.09 and ROC AUC of 0.91 ± 0.06. The detailed precision, recall, and F1 scores for each class and average are reported in Figure 6 (b). In Figure 6 (a), we present the average ROC curve with standard deviations.

Next, since our 12 features are explainable, we examined the feature importance from the RF classifier, as shown



(a) ROC curves for Random Forests to classify individuals with anxiety. We carried out a stratified 5-fold cross-validation. The grey area represents ± 1 standard deviation.

5-fold cross validation	Not anxious	Anxious	Avg.
Precision	0.81 (± 0.12)	0.86 (± 0.07)	0.84 (± 0.10)
Recall	0.80 (± 0.11)	0.87 (± 0.09)	0.84 (± 0.10)
F1 score	0.80 (± 0.10)	0.86 (± 0.07)	0.83 (± 0.09)

(b) The performance of RF on the anxiety classification task. We carried out a stratified 5-fold cross-validation. The values after the ± sign represent 1 standard deviation

Figure 6: Performance of the RF model in classifying individuals with anxiety conditions.

Feature	Importance	Feature	Importance
$C_H^{weekday}$	0.12	α	0.09
$C_H^{weekend}$	0.11	β	0.10
C_H^{total}	0.01	γ	0.14
$T_H^{weekday}$	0.11	$\mathcal{I}_8$	0.10
$T_H^{weekend}$	0.10	$\mathcal{I}_9$	0.09
T_H^{total}	0.01	$\mathcal{I}_{10}$	0.02

Table 1: Feature importance from the RF classifier.

in Table 1. We observed that the feature weights are significantly lower for C_H^{total}, T_H^{total}, and $\mathcal{I}_{10}$. Other features shared moderately even importance. Total time and category entropy weighted less in the RF model. It may imply that user online behaviors and habits during weekdays are fairly distinct from that during weekends.

Predicting Anxiety for Individuals

In this section, we consider the anxiety score prediction task: **given the online data and the past anxiety level of an individual, we aim to estimate the future GAD-7 score for that individual.** Concretely, given the two rounds of data, we aim to predict the GAD-7 score in the *follow-up* round given the online history data *and* the GAD-7 score from the *first-round* of an individual. Formally, this task is regarded as a regression problem, and we are interested in $P(y_2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2, y_1)$.

This setup can be used as a guideline to initiate specific treatment steps. Counselors can use the model on a weekly basis to monitor anxiety levels of their patients remotely *in-between* sessions/follow-up visits. It enables caregivers to note abnormal spikes in the estimated level of anxiety *comparing* to the last visit. Healthcare providers can then either schedule an immediate follow-up or use this information to engage with the patient to uncover issues that may otherwise go unmentioned during the next appointment.

For predicting anxiety scores, y_2, we only consider the significant features from the RF classifier above: weekday/weekend Time & Category entropy $\{C_H^{weekday}, C_H^{weekend}, T_H^{weekday}, T_H^{weekend}\}$, the Temporality parameters $\{\gamma, \alpha, \beta\}$, and the Inactivity Periods with thresholds of 8 and 9 hours $\{\mathcal{I}_8, \mathcal{I}_9\}$ as inputs. Thus, for the rest of the section, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in R^9$ for all individuals.

We hypothesize that the change in online behaviors may preserve information about the change in anxiety level. To leverage this in the prediction task, we define the following

feature vectors for the regression models:

$$\Delta \boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2 \in \mathbb{R}^9 \tag{3}$$

$$\boldsymbol{x}_{gp} = [\eta \odot \boldsymbol{x}_2, (1-\eta) \odot \Delta \boldsymbol{x}] \in \mathbb{R}^{2\times 9} \tag{4}$$

$$\boldsymbol{x}_{reg} = [\underbrace{\eta \odot \boldsymbol{x}_2, (1-\eta) \odot \Delta \boldsymbol{x}}_{\boldsymbol{x}_{gp}}, y_1] \in \mathbb{R}^{2\times 9+1} \tag{5}$$

where the square bracket indicates concatenation, $\eta \in [0, 1]$ is a hyperparameter that controls the weight on $\boldsymbol{x}_2$ and $\Delta \boldsymbol{x}$, and $\odot$ denotes an element-wise multiplication. $\boldsymbol{x}_{gp}$ is a trivial modification of $\boldsymbol{x}_{reg}$ by slicing out the last entry y_1 and keeping only the online data features. The intuition is that $\Delta \boldsymbol{x}$ captures the shift in online behaviors between two rounds; $\boldsymbol{x}_2$ is the most recent online observation in predicting y_2; y_1 acts as a base point of y_2; η weights the importance between $\Delta \boldsymbol{x}$ and $\boldsymbol{x}_2$.

We chose $\eta = 0.9$ and fed the $\boldsymbol{x}_{reg}$ as inputs. We first trained an Ordinary Least Squares regression (OLS). It achieved an average mean square error (MSE) of 4.77 ± 0.25 in predicting future GAD-7 scores y_2 (see Table 2).

Instead of merely looking for the best prediction given by maximum likelihood estimations, it is crucial to assess the uncertainty over the model and take a Bayesian perspective, especially given we are working with healthcare applications with limited sample size. Moreover, it would grant much flexibility if the regression is not limited to parametric linear form but in a functional space with non-linearity, investigating the distribution of *functions*. Therefore, we performed the regression task with a non-parametric Bayesian method, the Gaussian Process (GP) (Williams and Rasmussen 2006). We define our regression function as $f(\boldsymbol{x}_{reg})$, and it follows the GP below:

$$f(\boldsymbol{x}_{reg}) \sim \mathcal{GP}\left(m(\boldsymbol{x}_{reg}), k(\boldsymbol{x}_{reg}, \boldsymbol{x}'_{reg})\right) \tag{6}$$

$$m(\boldsymbol{x}_{reg}) = y_1 \tag{7}$$

$$k(\boldsymbol{x}_{reg}, \boldsymbol{x}'_{reg}) = exp\left(-\frac{\|\boldsymbol{x}_{gp} - \boldsymbol{x}'_{gp}\|^2}{2\ell}\right) \tag{8}$$

$$y_2 = f(\boldsymbol{x}_{reg}) + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma) \tag{9}$$

where $m(\boldsymbol{x}_{reg})$, the mean of the GP, is a deterministic function that returns the corresponding previous anxiety score y_1 for each subject. The covariance matrix is obtained by an exponential quadratic kernel k over all pairs of individual online data, $(\boldsymbol{x}_{gp}, \boldsymbol{x}'_{gp})$. It entails that, given any pair of individuals, the closer the distance between their online activity features in the vector space, the greater the correlation between their anxiety scores y_2 (close to 1), and *vice versa* (close to 0). ℓ is a hyperparameter that controls the length scale between data points: the greater the ℓ, the smoother the function. We further assume that the true y_2 equals to the function prediction plus an independent unknown Gaussian noise ϵ, and σ is the hyperparameter for the noise distribution. The above GP gave us a prior belief over the possible regression functions. The intuition is that, in the output space of our function $f(\boldsymbol{x}_{reg})$, the future GAD-7 anxiety scores, y_2, are normally distributed with a mean of the previous anxiety scores, y_1. The correlations between different y_2 values are determined by the similarities between online activities $\boldsymbol{x}_{gp}$ from the input space.

Model	(a)	(b)
OLS	4.77 ± 0.25	4.93 ± 0.32
GP	$\mathbf{1.87 \pm 0.14}$	$\mathbf{1.86 \pm 0.15}$

Table 2: The MSE of OLS and GP in predicting GAD-7. Column (a) shows the normal cross-validation. Column (b) shows the performance on the 13 subjects with significant anxiety changes.

In order to assess the performance of our GP over the test set, we first obtained the predictive posterior:

$$P\left(f\left(\boldsymbol{x}_{reg}^{test}\right) \mid f\left(\boldsymbol{x}_{reg}^{train}\right), \boldsymbol{x}_{reg}^{train}, \boldsymbol{x}_{reg}^{test}\right) \tag{10}$$

over all the regression functions conditioned on (after observing) the training set. After that, we sampled 100 functions (traces) from the posterior in Equation 10 and used them to make predictions on the test set. We report the average MSE of the 100 functions. Such process is repeated for each fold in the cross-validation. We report the average performance over the 5 folds in Figure 2. Our GP achieved an average MSE of 1.87 ± 0.14 in predicting future anxiety scores y_2.

There were 13 individuals whose ground truth GAD-7 anxiety scores changed by more than 5 between the *first-round* and the *follow-up*. A change in 5 of GAD-7 scores (ranging from 0 to 21) represents a change in anxiety level by around 23%, which can be clinically alarming. Thus, we conducted another 5-fold cross-validation but kept all these 13 subjects in the test set of each fold. We observed a good flexibility of $\boldsymbol{x}_{reg}$ in capturing such significant changes in GAD-7 since the performances are comparable to the average scores for all models, see Table 2, column (b).

Discussion

In this paper, we ran a novel longitudinal study that collected ubiquitous online activities logs along with gold-standard clinically validated anxiety scores. We have developed explainable features that capture various semantic and temporal facets of online engagement logs, such as activity and inactivity patterns, content and time diversities. We have shown that these features are strong signals for not only detecting individuals with anxiety disorders but also estimating the severity of anxiety given any segment of online activity history. Given one's online activities, our best performing classifier can reach an average F1 score of 0.83, average precision and recall of 0.84, and average AUC of 0.90. Furthermore, we have demonstrated that anxiety scores can be predicted with high accuracy with an average MSE of 1.87. To the best of our knowledge, we are the first to study and demonstrate that it is feasible to identify whether one is experiencing anxiety and estimate his/her/their exact anxiety score using individual-level YouTube and Google Search engine history logs. Our findings suggest the viability of constructing remote mental health surveillance frameworks based on passively sensed online data, which may be cheap, efficient, and bypasses the patient reluctance and information concealing dilemmas of traditional systems.

The Curse of Variability: Inferring mental health conditions such as anxiety from online behavior is challenging due to the wide array of subjective and external factors, such as seasonality, environment, etc., that add questionable variability in assessing one's level of anxiety. For example, some individuals may choose to not use any online platforms while experiencing anxiety. Someone may be very concerned about his/her/their significant other's anxiety disorder and research on the web, which may result in a false positive image. Furthermore, one may not be near a computer or mobile device when he/she/they are experiencing anxiety, and hence a framework such as ours may miss out on capturing signals that may be associated with anxiety. Besides, how people conduct searches on YouTube and Google Search engine is subjected to change over time. One possible way to address such high variability is to incorporate longitudinal studies on large populations. However, such studies require time and can be expensive.

The Prevalence of Uncertainty: We acknowledge that any mental health sensing system, such as our anxiety assessment framework, even under the most ideal circumstances, will likely have some degree of error and uncertainty. The trade-off between accuracy and uncertainty should be considered prior to designing a mental health sensing system. Lim et al. and Kay et al. have explored questions around how much uncertainty is acceptable, how much accuracy is sufficient, and how to best mitigate the uncertainty (Lim and Dey 2011; Kay, Patel, and Kientz 2015). There are open questions such as the cost of misclassification, how derived models around mental health indicators can be integrated in the current system need more attentions. A clear guideline needs to be set through discussions among therapists, clinicians, and computer scientists. Furthermore, models derived from such a specific population (college population) may not generalize at population level. We acknowledge that further investigation is required on a diverse population.

Privacy & Ethical Considerations: Building an anxiety monitoring system using individual-level YouTube and Google Search engine activity logs presents a series of concerns around privacy and data safety. Due to the sensitive nature of the data collected in this study, it is important that appropriate human subject protection protocols are in place. Hence our HIPAA-compliant study protocol has been rigorously reviewed and approved by our Institutional Review Board to address these concerns. Despite these measures, we acknowledge that ethical challenges may still arise if applications based on our methods are deployed in the real world.

When someone uses platforms such as YouTube and Google Search engine, he/she/they never intend the personal data to be used by mental health assessment systems. Hence, some individuals may choose not to share their sensitive data and refuse to participate. It is important to ensure that participants, at all times, have the choice and control over their data and can choose to exclude themselves from such studies at will. Participants need to be explicitly informed about how their online engagement logs will be de-identified and analyzed, what type of information it may reveal about the user, and the accrued benefits to the patients and the therapists/care providers from mental health clinics. To address these concerns, we employed an opt-in model for volunteering study participation. In addition, we conducted one-on-one interviews for each participant during the recruitment procedure so that the research team can (a) take the time to clearly explain the purpose and the outcome of the study and (b) explicitly inform the participants about the existence of such sensitive data and how they reserve full control over the information shared such as limiting data access or deleting data. Yet, one big limitation of employing opt-in model is that it may significantly limit the number of volunteering participants for the study. Besides, the opt-in procedure may introduce participation bias in terms of study recruitment and the awareness of subjects. To limit recruitment bias, we have adapted generic wordings, such as "*help us learn about mental health using online data,*"in our study advertisements without specifically mentioning anxiety.

References

Adler, N.; Cattuto, C.; Kalimeri, K.; Paolotti, D.; Tizzoni, M.; Verhulst, S.; Yom-Tov, E.; and Young, A. 2019. How search engine data enhance the understanding of determinants of suicide in India and inform prevention: observational study. *Journal of medical Internet research* 21(1): e10179.

Association., A. C. H. 2018. American College Health Association-National College Health Assessment II: Undergraduate Student Reference Group Data Report Fall 2018. *Silver Spring, MD: American College Health Association* .

Ayers, J. W.; Althouse, B. M.; Allem, J.-P.; Rosenquist, J. N.; and Ford, D. E. 2013. Seasonality in seeking mental health information on Google. *American journal of preventive medicine* 44(5): 520–525.

Bagroy, S.; Kumaraguru, P.; and De Choudhury, M. 2017. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human factors in Computing Systems*, 1634–1646.

Calling, S.; Midlöv, P.; Johansson, S.-E.; Sundquist, K.; and Sundquist, J. 2017. Longitudinal trends in self-reported anxiety. Effects of age and birth cohort during 25 years. *BMC psychiatry* 17(1): 119.

Cayari, C. 2011. The YouTube Effect: How YouTube Has Provided New Ways to Consume, Create, and Share Music. *International Journal of Education & the Arts* 12(6): n6.

Chancellor, S.; and De Choudhury, M. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3(1): 1–11.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 31–39.

Costello, E. J.; Egger, H. L.; and Angold, A. 2005. The developmental epidemiology of anxiety disorders: phenomenology, prevalence, and comorbidity. *Child and Adolescent Psychiatric Clinics* 14(4): 631–648.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56. ACM.

De Choudhury, M.; Counts, S.; Horvitz, E. J.; and Hoff, A. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 626–638.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression via Social Media. In *ICWSM*, 2.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110. ACM.

Fleck, B. K.; Beckman, L. M.; Sterns, J. L.; and Hussey, H. D. 2014. YouTube in the classroom: Helpful tips and student perceptions. *Journal of Effective Teaching* 14(3): 21–37.

Gil-Or1, O.; Levi-Belzm, Y.; and Turel, O. 2015. The "Facebook-self": characteristics and psychological predictors of false self-presentation on Facebook. *Frontiers in Psychology* 6: 99.

Google. 2020. Content Categories. https://cloud.google.com/natural-language/docs/categories. [Online; accessed 12-May-2020].

Gopalakrishna Pillai, R.; Thelwall, M.; and Orasan, C. 2018. Detection of stress and relaxation magnitudes for tweets. In *Companion Proceedings of the The Web Conference 2018*, 1677–1684.

Gunn III, J. F.; and Lester, D. 2013. Using google searches on the internet to monitor suicidal behavior. *Journal of affective disorders* 148(2): 411–412.

Hall, A. S.; and Parsons, J. 2001. Internet addiction: College student case study using best practices in cognitive behavior therapy. *Journal of mental health counseling* 23(4): 312.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58(1): 83–90.

Jamison-Powell, S.; Linehan, C.; Daley, L.; Garbett, A.; and Lawson, S. 2012. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1501–1510. ACM.

Jimenez, A.; Santed-Germán, M.-A.; and Ramos, V. 2020. Google searches and suicide rates in Spain, 2004-2013: correlation study. *JMIR public health and surveillance* 6(2): e10919.

Kaplan, A. M.; and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53(1): 59–68.

Kay, M.; Patel, S. N.; and Kientz, J. A. 2015. How good is 85%? A survey tool to connect classifier evaluation to acceptability of accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 347–356.

Kiang, A.; and Bailon, J. 2016. Data loss prevention (DLP) methods and architectures by a cloud service. US Patent 9,237,170.

Kim, T. W.; and Paek, S. T. 2016. Cloud data discovery method and system for private information protection and data loss prevention in enterprise cloud service environment. US Patent App. 14/728,503.

Klobas, J. E.; McGill, T. J.; Moghavvemi, S.; and Paramanathan, T. 2018. Compulsive YouTube usage: A comparison of use motivation and personality effects. *Computers in Human Behavior* 87: 129–139.

Lim, B. Y.; and Dey, A. K. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing*, 415–424.

McCarthy, M. J. 2010. Internet monitoring of suicide risk in the population. *Journal of affective disorders* 122(3): 277–279.

Moghavvemi, S.; Sulaiman, A. B.; Jaafar, N. I. B.; and Kasem, N. 2017. Facebook and YouTube addiction: the usage pattern of Malaysian students. In *2017 international conference on research and innovation in information systems (ICRIIS)*, 1–6. IEEE.

Mohr, D. C.; Zhang, M.; and Schueller, S. M. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13: 23–47.

Nuti, S. V.; Wayda, B.; Ranasinghe, I.; Wang, S.; Dreyer, R. P.; Chen, S. I.; and Murugiah, K. 2014. The use of google trends in health care research: a systematic review. *PloS one* 9(10): e109583.

Ophir, Y.; Asterhan, C.; and Schwarz, B. 2020. If these Facebook walls could talk: Detecting and treating teenage psycho-social stress through social network activity (in Hebrew). *Breaking down barriers? Teachers, students and social network sites* .

Paparrizos, J.; White, R. W.; and Horvitz, E. 2016a. Detecting devastating diseases in search logs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 559–568. ACM.

Paparrizos, J.; White, R. W.; and Horvitz, E. 2016b. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice* 12(8): 737–744.

Purdon, C.; Antony, M.; Monteiro, S.; and Swinson, R. P. 2001. Social anxiety in college students. *Journal of Anxiety Disorders* 15(3): 203–215.

Reece, A. G.; Reagan, A. J.; Lix, K. L.; Dodds, P. S.; Danforth, C. M.; and Langer, E. J. 2017. Forecasting the onset

and course of mental illness with Twitter data. *Scientific reports* 7(1): 1–11.

Rizoiu, M.-A.; Lee, Y.; Mishra, S.; and Xie, L. 2017. Hawkes processes for events in social media. In *Frontiers of Multimedia Research*, 191–218.

Schwartz, H. A.; Eichstaedt, J.; Kern, M.; Park, G.; Sap, M.; Stillwell, D.; Kosinski, M.; and Ungar, L. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 118–125.

Seabrook, E. M.; Kern, M. L.; and Rickard, N. S. 2016. Social networking sites, depression, and anxiety: a systematic review. *JMIR mental health* 3(4): e50.

Shenkin, P. S.; Erman, B.; and Mastrandrea, L. D. 1991. Information-theoretical entropy as a measure of sequence variability. *Proteins: Structure, Function, and Bioinformatics* 11(4): 297–313.

Spitzer, R. L.; Kroenke, K.; Williams, J. B.; and Löwe, B. 2006. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine* 166(10): 1092–1097.

Sueki, H. 2011. Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. *Psychiatry and clinical neurosciences* 65(4): 392–394.

Swinson, R. 2006. The GAD-7 scale was accurate for diagnosing generalised anxiety disorder. *Evidence-based medicine* 11(6): 184.

TechPostPlus. 2019. YouTube video Categories list FAQs and solutions. https://techpostplus.com/2019/04/26/youtube-video-categories-list-faqs-and-solutions/. [Online; accessed 26-April-2019].

The Next Web. 2020. Digital trends 2020: Every single stat you need to know about the internet. https://thenextweb.com/podium/2020/01/30/digital-trends-2020-every-single-stat-you-need-to-know-about-the-internet. Accessed: 2020-02-07.

Whiteford, H. A.; Degenhardt, L.; Rehm, J.; Baxter, A. J.; Ferrari, A. J.; Erskine, H. E.; Charlson, F. J.; Norman, R. E.; Flaxman, A. D.; Johns, N.; et al. 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The lancet* 382(9904): 1575–1586.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Williams, N. 2014. The GAD-7 questionnaire. *Occupational medicine* 64(3): 224–224.

Yang, A. C.; Tsai, S.-J.; Huang, N. E.; and Peng, C.-K. 2011. Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009. *Journal of affective disorders* 132(1): 179–184.

Zaman, A.; Acharyya, R.; Kautz, H.; and Silenzio, V. 2019. Detecting Low Self-Esteem in Youths from Web Search Data. In *The World Wide Web Conference*, 2270–2280.