# Twitter-STMHD: An Extensive User-Level Database of Multiple Mental Health Disorders

**Suhavi,[1] Asmit Kumar Singh,[*1] Udit Arora,[*1] Somyadeep Shrivastava,[*2] Aryaveer Singh,[*3] Rajiv Ratn Shah,[1] Ponnurangam Kumaraguru[4]**

[1]Indraprastha Institute of Information and Technology, Delhi, India
[2]Indian Institute of Information Technology, Dharwad, India
[3]Guru Gobind Singh Indraprastha University, Delhi, India
[4]International Institute of Information and Technology, Hyderabad, India
suhavi16099@iiitd.ac.in, asmit18025@iiitd.ac.in, udit18417@iiitd.ac.in, 17bcs028@iiitdwd.ac.in, aryaveersingh@acm.org, rajivratn@iiitd.ac.in, pk.guru@iiit.ac.in

## Abstract

Social Media is equipped with the ability to track and quantify user behavior, establishing it as an appropriate resource for mental health studies. However, previous efforts in the area have been limited by the lack of data and contextually relevant information. There is a need for large-scale, well-labeled mental health datasets with fast reproducible methods to facilitate their heuristic growth. In this paper, we cater to this need by building the Twitter - Self-Reported Temporally-Contextual Mental Health Diagnosis Dataset (Twitter-STMHD), a large scale, user-level dataset grouped into 8 disorder categories and a companion class of control users. The dataset is 60% hand-annotated, which lead to the creation of high-precision self-reported diagnosis report patterns, used for the construction of the rest of the dataset. The dataset, instead of being a corpus of tweets, is a collection of user-profiles of those suffering from mental health disorders to provide a holistic view of the problem statement. By leveraging temporal information, the data for a given profile in the dataset has been collected for disease prevalence periods: onset of disorder, diagnosis and progression, along with a fourth period: COVID-19. This is the only and the largest dataset that captures the tweeting activity of users suffering from mental health disorders during the COVID-19 period.

## Introduction

Depression. Anxiety. Bipolar disorder. Obsessive Behaviors. Trauma. These are a few commonly known mental health illnesses or disorders, which alter the sufferer's emotions, mood, thought, behavior, altering their entire lives, yet lack clear physical evidence that indicates their presence.

As of 2017, Information for Health Metrics and Evaluation in a survey estimated 792 million people globally lived with a mental disorder, accounting for over 10% of the world's population (Saloni Dattani and Roser 2021). Mental health disorders deter lives, alter interpersonal relations, drive down productivity rates, ultimately affecting countries' economies, prevailing as the top-ten contributors to the global health-related burden since 1990 (GDB 2021). World Health Organization states that in its worst form, mental health disorders can lead to suicide ideation (Organization 2019), the second leading cause of death among 15-29 year old over 2007-2017, emphasizing the importance of early diagnosis. Clinically, professional psychologists diagnose mental health disorders in face-to-face interviews, using DSM-V, The Diagnostic and Statistical Manual of Mental Disorders, published in 2013, as the reference. The DSM-IV, published in 1994, predecessor of DSM-V till 2013, (Kendler 2013) describing mental Health disorders notes (A., Frances, and APA 1994),

> ”Mental disorders have also been defined by a variety of concepts (e.g., distress, dyscontrol, disadvantage, disability, inflexibility, irrationality, syndromal pattern, etiology, and statistical deviation). Each is a useful indicator for a mental disorder, but none is equivalent to the concept, and different situations call for different definitions.”

In addition to the lack of any determinative symptom, several demographic factors add to the challenges to diagnosis of mental health disorders, like unawareness and lack of resources prevalent in mid to low income countries, longstanding social stigma making it a taboo, and imperfect recall of mood and behavioral changes over the observation period (mood and behavioral reports from patients are the basis of clinical diagnosis by psychologists). While the traditional methods are the most effective, they continue to be slow and time-consuming. 76% to 85% of people remain undiagnosed and untreated worldwide (James et al. 2018).

On the other hand, over the past decade, people have taken to social media platforms like Facebook, Twitter and Reddit to emote freely, interact with content daily, make and keep up with friends, share personal news. As of October 2021, an estimated 4.55 billion people (57.6% of the global population) used social media, out of which Twitter recorded 436.4 million users (DataReportal 2021). The COVID-19 outbreak forced people inside their homes, cutting them from physical

[*]These authors contributed equally.

worlds, escalating the usage of social media, making them resort solely to the platforms for socializing, emoting and interacting with people. A study by Singh, Dixit, and Joshi, analyzing the reason behind compulsive usage of social media during the pandemic called it a "psychological necessity" catering to people's needs for human interaction. As Social Media emerged as a popular tool for coping with the pandemic (WHO 2019), people's Mental Health under the "new normal" declined drastically (Singh, Dixit, and Joshi 2020) (Pfefferbaum and North 2020) (IHME 2021) (Santomauro et al. 2021). Santomauro et al. in their paper call for the need for an up-to-date, heuristically growing information database to deal with the problem effectively and promptly. User-Generated Content (posts, images, videos, replies, likes, upvotes, shares) (UGC) on social media instantly reflects users' daily lives and mental states. If leveraged correctly, social media can act as a resource for a precise, real-time, heuristically-growing database on mental-health and wellness studies which is the aim of our paper.

A dataset for mental health research should ideally contain: *For any given disorder:* a wide variety of users facilitating enough data to examine unbiased and generalised results. *For any given user (in disorder class):* at least the information taken into context in a typical disorder diagnosis interview conducted by a certified practitioner.[1] We maintained the above as the skeleton aim for our study.

The realization of social media as an important mental health resource gained popularity in the past decade, leading to novel analytical studies and datasets customized for the respective use cases (Coppersmith, Dredze, and Harman 2014a) (De Choudhury et al. 2013). Prior work relied on self-disclosure, either through self-opted questionnaires/surveys (De Choudhury et al. 2013), or via textual posts on social media platforms self-disclosing the diagnosis. The earliest attempts for identifying self-disclosure diagnosis statements used micro-blogging sites like Twitter (Coppersmith, Dredze, and Harman 2014a) and Facebook (Park et al. 2013) (De Choudhury et al. 2014) (Sap et al. 2014). The aim was to identify subtleties in language that could identify potential at-risk users. Once identified, the users can be made aware and provided with help and resources. Coppersmith et al. constructed a large dataset for various mental-health disorders selecting users via self-disclosed disorder diagnosis tweets. However, the criteria for user tweet collection still catered only to the idea of collecting enough text, skipping any temporal or social engagement data and its variations with time. The granular nature of a post-level dataset fails to present a complete picture and, in some cases, also misses the context required for a time-sensitive use-case like early detection. For example, a person self-reporting a diagnosis for depression in 2016 might not contain or exhibit any depression language in 2022. The assumption that the language characteristics identified from the tweets of 2022 can map the user's mental health status at the time of the diagnosis is invalid. The correct mental health data for identifying early-depression signs in the user must have data from before the self-reported diagnosis tweet to study the behaviors in the period that mark the onset of the disease. Temporal reference for tweets arms us with the capability to distinguish between the different periods of the users' journey of living with the disorder, like the onset of the disorder, the diagnosis, and disease progression, thus becoming a crucial attribute for potential at-risk users (MacAvaney et al. 2018). Temporal information, in addition to that, gives flexibility with the kind of use-cases that a given dataset can cater to. Shen et al. published a dataset of 3600 depressed users using Twitter. They selected users through self-disclosed diagnosis tweets, but instead of creating a corpus dataset of tweets, they created a dataset of user-profiles containing profile information, activity, interests, and timeline tweets along with the timestamps and metadata for each tweet, for a period of 3-4 months before the diagnosis, thus not treating depression as an adjective which can be assessed from a single tweet.

We extend from the previous efforts for addressing the need for labeled, large-scale, temporally contextual mental health datasets in our work. In particular, we improve upon collection methods for preparing high-precision datasets, bringing temporal context to user activities and focusing on user-level studies instead of contextless post-level studies. We have built the Twitter Self-reported Temporally-contextual Mental-Health User-Level Dataset, "STMHD", of users picked via self-disclosure tweets, grouped into eight mental-health disorder groups and a corresponding class for control users. The dataset is a collection of user profiles with tweeting activities from 3 broad disorder prevalence periods, onset, diagnosis and progression of disorder, to cater to use-cases like early detection of disease and identify at-risk users. Keeping in touch with the real world, we have added a fourth temporally relevant period, the COVID-19 period, to facilitate studies on understanding the stark negative effect of the pandemic on mental health. Our contributions to the mental-health research space are as follows:

- **A large-scale 8-class mental health dataset**, accompanied by a control-users class.
- **High precision, 60% hand-annotated dataset**, we weeded out the false-positives, self-disclosures that emerge because of loose regex patterns like "diagnosed with <disorder name>".
- **Lexicon for heuristic growth**, we built a lexicon of specialized patterns for accurately identifying self-reported diagnosis reports eliminating the need for manual annotation which was then used for constructing 40% of our database.
- **Temporal context** Realizing the need for a temporal context for user activities, we recorded three primary disorder prevalence periods for collecting user-timeline and profile data; onset, diagnosis and progression, as well as the COVID-19 period.
- **User-level database** We surpassed the tweet-level corpus nature of mental health datasets by collecting profiles instead, putting the complete user behavior into context, much similar to building a case history in a clinical diagnosis interview.

[1] The DSM-V contains descriptions, symptoms, and other criteria for diagnosing mental disorders, along with an approximate period of symptom and disease prevalence.

- **Space for users with multiple disorders** Our dataset contains users with more than one disorder, thus allowing behavioral studies of interacting mental health conditions in line with real-world use cases.
- **Dataset validation and feature groups** We assembled the information into possible feature groups, quantifying the data and trends in data confirming findings in previously conducted research, thus validating our methods.
- **First, largest, Covid-19 mental health dataset** The pandemic period was noted for deterring mental health at large. We, thus, captured two kinds of users, those with new diagnosis reports during the pandemic and those with diagnosis reports from before, quantifying the implications of the pandemic on their mental health. Our dataset effectively captures how a global-level real-world change affected both, the control users and disorder-category users via a change in their Twitter activities.

## Literature Review

The lack of well-labelled, large-scale mental health datasets has been a critical challenge to the research domain since the beginning despite the ubiquity of social media. *Social* Media, as the name itself indicates, first attracted the attention of researchers as a possible mental resource as it provided a medium for interacting with people (being social) using platform features like sharing posts, liking, down-voting, and commenting. Mental health conditions are primarily concerned with behavioral and mood changes and social media became a medium that captured these aspects for a user in a quantifiable way. The connection between the two is indisputable. Coppersmith, Dredze, and Harman remarks that while social media had previously been used for several diseases, using it for gaining insight into mental health is perhaps the most appropriate use out of all. The earliest studies scanned Twitter for depression-related discourse over two months to check its validity as a mental-health resource; successful observations led to a dataset of tweets from 69 users(Park, Cha, and Cha 2012). Following works continued to rely on outside social-media information like disclosure forms for user identification and personal interviews for behavioral data. (De Choudhury et al. 2013) (Park et al. 2013) (Wang et al. 2013).

The high cost and bias of relying on forms shifted the identification process to depend on self-reported diagnosis posts by social media users. A typical self-reported diagnosis tweet reads as *"I have been diagnosed with depression."* (Coppersmith, Dredze, and Harman 2014b) (Coppersmith et al. 2015c) (Coppersmith et al. 2015b) (Coppersmith et al. 2016). To identify post-partum depression, De Choudhury et al. created a dataset of new mothers leveraging baby-announcement posts on Twitter. The dataset contained an equal number of tweets from pre-natal (pre-childbirth) and post-natal (post-childbirth) periods. Their classification framework accuracy jumped from 71% to 80-83% by simply using the data from the postnatal period in addition to prenatal data, highlighting the importance of temporal context in mental-health datasets as the mental state of a person varies with time.

Coppersmith et al. constructed a large dataset covering 11 unique disorders; however, the dataset was a corpus of the last few tweets of every user, selected irrespective of the context of onset or progression of the disease. The tweets from all users were taken together and granularly analyzed for linguistic patterns. The practice, by default, assumes the possibility of a disorder detection using a single post or tweet. This assumption contradicts the DSM-V mandated periods of symptom prevalence before a diagnosis can be made. Short lengths of textual data from Twitter, while being considered singularly, posed a low-context problem, attracting mental health datasets from the discussion-forum social platform, Reddit, which has no character limits on posts. Earlier efforts depended on mental-health or disease-related subreddit participation to identify users (Kumar et al. 2015) (Bagroy, Kumaraguru, and De Choudhury 2017). Cohan et al. used self-disclosure reports to build a large-scale Reddit user dataset of nine different disorders to facilitate an extensive linguistic study for the identified users and the control users.

Shen et al. used self-disclosed diagnosis tweets to build a user-level dataset of about 3600 depressed class users. The dataset contains two types of information for every user: profile statistics and timeline tweets (for a month before the diagnosis report tweet). Every individual tweet collected has its own temporal and engagement context. The study focused on feature groups beyond linguistics, like social engagement via followers-following count, user-networks via post engagement data, topic-level features via the kind of topics discussed in tweets and the trends were compared for negative-class and depressed-class users. Most notably, the dataset included timestamp data for all user activities to indicate sleeping patterns, using which as a feature group, the study devised a gold-standard classifier framework. The user-level granularity of this dataset resolved the issue of low character-cutoff limits on Twitter, as a large number of tweets could now be appended sequentially owing to tweet timestamp data with or without temporal weights.

Our work identifies the limitations of the previous datasets, specifically improving upon Shen et al.'s work on user identification and data collection for users. Twitter-STMHD was constructed to provide a more holistic view of users' information, much similar to building a case-study in traditional diagnosis practices. The dataset aims to cater to the need for a well-labeled, temporally-relevant large-scale mental health dataset to aid research in the domain.

## Data

Twitter-STMHD contains eight mental health disorder classes, each corresponding to branches in the DSM-V. Three of the studied conditions, namely depression, major-depressive disorder (mdd) and post-partum depression (ppd), belong to the class of depressive disorders; post-traumatic stress disorder (ptsd) is a subset of trauma and stress-related disorders, and attention-deficit/hyperactivity disorders (adhd) belongs to neurodevelopment disorder class. Others are anxiety disorders (anxiety), bipolar disorders (bipolar) and obsessive-compulsive disorders (ocd). A ninth class, the control-users class, is released as part of the dataset. The dataset contains 25,860 unique users belonging to at least one of the eight disorder categories and approximately 8000 control users. [2]

## Dataset Structure and Collection Period

The T-STMHD was created by selecting users with self-reported diagnosis disclosure posts on Twitter. We will refer to such posts (tweets) in which a user claims to have been diagnosed with one of the eight mental health conditions as the user's anchor tweet for the purpose they serve and ease of reference. The control class contains users least likely to have any of the disorders discussed in this study. For each user, the dataset provides two types of information, user-profile data and timeline-tweets data, containing all tweets from the user-specific disorder prevalence period. The approach is based on the dataset construction methods used by Shen et al.

Every user-profile was assigned a unique timeline using its anchor tweet timestamp; this defined its data collection period. The length of the period taken into account was kept uniform across all users of all disorder classes. The range of the collection window must cover the period that potentially contains the onset of the condition before the diagnosis and the disease progression and prevalence period post the diagnosis. We took the maximum of all observation periods suggested in the DSM-5 for the eight disorders to assign a uniform length to the collection period. $\Delta =$ max($T_{depression}$, $T_{anxiety}$, $T_{ocd}$, $T_{bpd}$, $T_{ptsd}$, $T_{mdd}$, $T_{ppd}$, $T_{adhd}$) = 2 years

Consequently, a 4-year window was chosen, spanning two years before the anchor tweet to two years after it. Readily available and attribute-wise flexible Twitter APIs were used for data extraction, which allowed us to shape our data into the required structure.

## COVID-19 Period

COVID-19 depleted people's mental health while simultaneously recording a surge in social media activity. We thus extended the dataset to include another temporally contextual period; COVID-19, by including two kinds of user groups in each class; users with anchor tweets from before the pandemic; January 2017 to March 2020 and a new set of users with anchor tweets dated post the announcement of the pandemic; April 2020 - May 2021. For the first kind of user-proflies, we collected their Twitter activity during the pandemic in addition to the 4-year disease prevalence window. For the second kind, the pandemic period was by default a subset of the data collection window.

## Anchor Tweet Identification

To identify users, we needed to identify the anchor tweets correctly.

**Preliminary set of anchor tweets** The first step to identifying anchor tweets was to get all tweets containing a loose pattern which was passed as a query to the API. A typical anchor tweet has two parts: a self-disclosure of diagnosis and the disorder's name. We used the loose pattern: "diagnosed with <disorder name>", where the word 'diagnosed' suggests a clinical diagnosis. For disorder names, we prepared a lexicon using common synonyms, formal DSM-V names, mis-spellings and abbreviations to capture the various possible probable anchor tweets. While using a loose pattern gave us all tweets containing the pattern, it also led to several false positives. Typical examples are 'My mother got diagnosed with adhd' and 'I was not diagnosed with depression'. Hence the need for methods to eliminate the false anchor tweets for an accurate dataset. This corpus of tweets became our preliminary set of anchor tweets.The number of collected tweets per disorder is present in Table 1. We followed two approaches for eliminating incorrect anchor tweets. We divided the set into two equal parts for each disorder, one part was hand-annotated, and the other matched against high-precision patterns.

**Hand annotation** We manually annotated the base corpus to eliminate false instances of anchor tweets. The annotation process required annotators to go over the tweets one-by-one. If the tweet indicated that its author had been diagnosed with the disorder class that that tweet was categorized into, we marked it as *positive*. We marked the tweets that indicated anything else as *negative*. Data was hand-annotated by five contributors. Each disorder class's preliminary corpus was divided into five equal parts and assigned to the contributors for annotation. We divided each part further into four equal parts and assigned each subpart to the other four annotators, who then annotated it again, irrespective of the previous annotation. All tweets, thus, were annotated twice. If both the annotations on a given tweet corresponded to a *positive*, we marked it as a valid anchor tweet. We annotated around 76000 tweets, out of which around 26000 were identified as true positives. These contribute to 60% of the users in our dataset. Table1 gives a disorder wise breakup of the same. We took the help of a licensed clinical psychologist [3] to validate our annotation process. A sample of 500 tweets was randomly selected with the same ratio of tweets belonging to each disorder class as in the original dataset. The psychologist was asked to tag each tweet as either a valid or an invalid anchor tweet.The annotations made by the psychologist were compared with annotations made by us. She disagreed with 4 out of the 500 annotations.Thus making 60% of our dataset 99.2% precise. [4]

**High-precision anchor tweet patterns** While hand-annotation accounts for the most precise and reliable datasets, it is a time-expensive procedure. The mental health research community needs to capture information and identify users as close to real-time as possible to avoid losing time-sensitive contexts in data that will ultimately be lost as subtleties in future papers studying historical corpora. Thus the need for high-precision patterns to identify valid anchor tweets.We studied the positively annotated tweets to prepare

[2] Dataset hosted at https://zenodo.org/record/6409736

[3] Dr. Shefali Gupta, M.Phil, RCI CRR No. A50454, Clinical Psychologist, Assistant Professor, Amity University, Gwalior

[4] The annotation sheets have been documented and added here: https://github.com/Suhavi/TrackingMentalHealth

| Disorders | Collected Tweets | Hand Annotated | Pattern Annotated | Final Anchor Tweets | Unfiltered User Count | Final User Count |
|---|---|---|---|---|---|---|
| **ADHD** | 43764 | 5039 | 3649 | 8688 | 8688 | 8095 |
| **Depression** | 37149 | 6791 | 4342 | 11133 | 11133 | 6803 |
| **PTSD** | 30077 | 3155 | 1854 | 5009 | 5009 | 3414 |
| **Anxiety** | 267339 | 5985 | 3669 | 9654 | 9654 | 4843 |
| **OCD** | 7558 | 1415 | 905 | 2320 | 2320 | 1325 |
| **PPD** | 713 | 333 | 263 | 596 | 596 | 547 |
| **MDD** | 651 | 331 | 179 | 510 | 510 | 325 |
| **Bipolar** | 5967 | 3168 | 2391 | 5559 | 5559 | 1651 |
| **Total Counts** | 152618 | 26117 | 17152 | 43269 | 43269 | 27003 |

Table 1: Represents the vast number of user anchor tweets first identified using a loose regex, against the number of valid ones recognised by hand-annotations and pattern matching. the last column lists the final user count for each disorders class in the dataset.

an elaborate lexicon containing the two identified parts of an anchor tweet and implemented a set of rules on their placement in the input text, upon parameters like the gap between the two, presence of negation words and the length of the text. We evaluated the performance of the hence created patterns on the hand-annotated tweets and were able to identify the positive users with a 94% precision, considering the hand annotations to be correct. These high precision string patterns were then used to identify anchor tweets from the other half of the preliminary database of anchor tweets, contributing to about 40% of the users in our dataset. Table 1 lists a disorder-wise distribution.

## Users and User-Data

**Timeline startpoint and endpoint** From the identified anchor tweets, the Twitter IDs of the users were extracted to make our preliminary database of users. We collected the Twitter activity of each user in this set between the dates $T_1$(start point)and $T_2$(endpoint), calculated as follows:

$$T_1 = T_{anchor} - \Delta \tag{1}$$

$$T_2 = max(T_{anchor} + \Delta, May2021) \tag{2}$$

Based on the afore determined collection period, we took a $\Delta = 2years$ before and after the self-diagnosis tweet timestamp to cover the several stages of disease prevalence. In addition, we collected the data for the COVID-19 dominant period, March 2020 to May 2021.

**Weeding out undesirable user accounts** To ensure that our dataset captures the general users on the platform who engage in natural and organic conversations we weeded out user profiles from the preliminary dataset based on two factors: Minimum tweet count; We removed users with less than 50 tweets in their data collection window to ensure that enough contextual data is present for each data point in our dataset as necessary for generalizing the results obtained from experimental studies. Maximum follower count; We removed any user with a follower count more than 5000 to ensure our dataset does not contain Twitter accounts of famous personalities that use Twitter as a brand advertisement tool or those of mental health and wellness organizations that use Twitter as a medium to indulge in wellness discourse. The final count of user-profiles in all 8-disorder classes is shown in Table 1.

**Control-user class** We assist the disorder classes with a class of control users. The number of users in this class is kept equal to the depression dataset class which has the highest number of user-profiles among the disorder classes. To ensure temporal consistency, we collected tweets randomly, sampled in the same range as the positive class; between January 2017 and May 2021. For each tweet, we recorded the user-id and collected profile data starting from two years prior to the posting time of the tweet till May 2021. For users to be least likely to belong to the eight disorder sets,we removed all those with any instance of indulgence in mental health discourse in their collected tweets. (Shen et al. 2017) We prepared a set of lexicons pertinent to mental health to carry out this task. Control users, too, were weeded out based on a minimum tweet, 50 count and a maximum follower, 5000 count to keep only the desirable, high-quality user profiles.

**User data collected** We collected two kinds of data for every user: the user profile information and the timeline tweets. Each of these has their own set of attributes. While collecting attributes, those with personally identifiable information such as profile name or links to the profile were removed to ensure that the user's identity is not compromised. The user-profile information contains the following attributes for each user: *creation_timestamp:* profile's date and time of creation, *description:* an intro input by the user displayed on the top of the profile, *favorites_count:* number of likes given by the user on Twitter, *friends_count:* number of mutual contacts on the profile, *follower_count:* number of profiles following the profile, *banner_link:* URL to a downloadable link of the banner picture, *display_image_link:* URL to a downloadable link of the display image, *status_count:* number of tweets posted by the user profile and *verified_check:* flag to note if user's account is verified or not. The timeline tweets file contains the textual and contextual information of all tweets made by the user in chronological order, starting from $T_1$ to $T_2$. The attributes collected for every tweet are: *text:* tweet content, *conversation_id:* the unique identifier of the Twitter thread the tweet is a part of, *tweet_id:* tweet's unique identifier, *language:* tweet language, *likes_count:* number of likes on the tweet, *quote_count:* number of times the tweet was quoted, *reply_count:* number of replies on the tweet, *retweet_count:* number of times the tweet was retweeted, *source_name:* tweet source, *timestamp_tweet:* tweet's time

| Disorder | Followers | Friends | Favourites | Status Count | Verified Percent |
|---|---|---|---|---|---|
| **ADHD** | 647.79 ± 875.29 | 806.78 ± 921.86 | 38871.45 ± 62133.02 | 19059.58 ± 32748.18 | 0.28 ± 0.28 |
| **Anxiety** | 715.85 ± 917.79 | 799.83 ± 941.28 | 33533.34 ± 52264.49 | 20478.89 ± 32218.89 | 0.21 ± 0.21 |
| **Bipolar** | 757.17 ± 980.56 | 895.66 ± 1058.72 | 27994.92 ± 49053.94 | 20500.12 ± 33302.43 | 0.24 ± 0.24 |
| **Depression** | 731.9 ± 943.06 | 740.51 ± 880.63 | 31265.83 ± 52077.97 | 23177.24 ± 38983.55 | 0.35 ± 0.35 |
| **MDD** | 782.08 ± 1007.67 | 750.06 ± 928.15 | 34309.69 ± 59937.54 | 27774.81 ± 57321.12 | 0.31 ± 0.31 |
| **OCD** | 707.95 ± 936.13 | 774.36 ± 936.3 | 35061.81 ± 57381.9 | 19180.28 ± 34700.96 | 0.15 ± 0.15 |
| **PPD** | 818.17 ± 1003.77 | 844.04 ± 996.72 | 21762.91 ± 51186.72 | 21235.07 ± 37636.88 | 0.4 ± 0.4 |
| **PTSD** | 810.12 ± 1028.43 | 905.28 ± 1085.31 | 28831.85 ± 50197.69 | 21603.24 ± 43315.24 | 0.53 ± 0.53 |
| **Control** | 743.19 ± 974.16 | 794.33 ± 990.57 | 21683.15 ± 44735.88 | 25320.8 ± 66340.87 | 0.67 ± 0.67 |

Table 2: Collected users aggregate statistics

and date of creation, *mentionedUsers:* the list of user ids mentioned in the tweet, *media:* all images and videos in the tweet with their own set of attributes. The attributes of an image include: *url:* downloadable URL of the image, *type:* set as ”image”. The attributes of videos are: *thumbnail_url:* downloadable URL to the video thumbnail, *url:* downloadable URL to the video, *bitrate:* maximum bitrate from multiple variants of the video and *type:* set as ”video”.

## Additional Features

**Tagging mental health discourse in dataset** In addition to the attributes obtained directly using the API, we added an attribute called *disorder_flag* which specifies if the tweet's text contains mental health discourse, using the same lexicon that was built to remove users from the control-users class. The tweets with the value 'True' were not removed (Cohan et al. 2018) but flagged for the scope of study of mental health discourse in users.

**Scope for multimodality** To give this dataset the scope of multimodality, each tweet from the user timeline data has a 'media' category that lists the kind of media (image, GIF, video) attached to a tweet and mentions a downloadable link for the same. All profile traceable or cross-platform identity mapping links were removed and replaced with downloadable links to images and videos. The links stop working only if the tweet is taken down or hidden from public view, in which case, the media, by default, becomes unfit for use under ethical considerations.

**Statistics** Statistics for user-profile data can be found in Table 2 giving an average range of the network size for each class of users. Tweet-level statistics can be found in Table 3 which shows the number of tweets, retweets, likes made and media shared by an average user of a given class.

# Data Quality

Twitter is a social and micro-blogging platform which boasts around 192 million daily active users. As of April 2021, Twitter's global audience was composed of 38.5% of users aged between 25 and 34 years old, 21% users aged between 35 and 49, 24% users below 24 years and users aged 50 or above accounted for around 17% (Statista 2021). Twitter roughly has 34% female and 66% male (Lin 2020). It has the most number of users from the USA, around 77 million, followed by Japan, India, Brazil, UK and several other countries (Clement 2018). All of these points make Twitter a demographically rich medium to collect our data. The readily available Twitter APIs which put no constraint on collecting data from any region or background enables our dataset to be more inclusive and diverse.

The average session on Twitter is 3.39 minutes with 500 million tweets sent out per day (Lin 2020). 78% of USA Twitter users like to express their opinions about topics they are knowledgeable about or interested in. Twitter users are considered highly influential, a Twitter-commissioned survey of friends of Twitter users in the UK found that 3 in 4 of them turn to Twitter for advice when they want to learn more about a topic (Stennis 2018). Henceforth, we can say that Twitter has quality users with active engagement. User relationships and activity is a determining factor in establishing social media dataset quality (Agichtein et al. 2008). Thus in order to get an overview of user tweeting activity and relationships for our dataset, we found out the average number of tweets made by users per year in a particular disorder class and an average number of followers and friends they have.

Figure 1 depicts that the control and disorder class users have similar distributions, which reflects upon the quality of user-profiles collected. The values aren't drastically low for any disorder class, which could have lead to the wrong hypothesis that users suffering from mental-health disorders

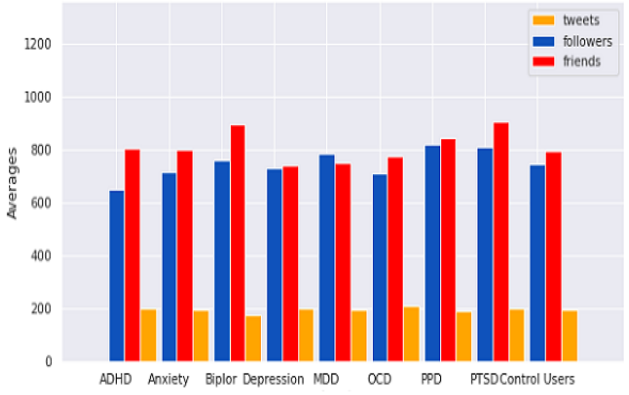

Figure 1: Users Relationship and Activity

| Disorder | Tweets Collected | Retweets | Likes | Replies | Mentioned Users | Media Count |
|---|---|---|---|---|---|---|
| **ADHD** | 7592.32 ± 13873.57 | 0.35 ± 2.32 | 3.19 ± 10.09 | 0.4 ± 0.22 | 0.79 ± 0.73 | 0.15 ± 0.13 |
| **Anxiety** | 10190.79 ± 16533.68 | 0.36 ± 3.02 | 2.56 ± 10.11 | 0.35 ± 0.25 | 0.74 ± 0.77 | 0.15 ± 0.14 |
| **Bipolar** | 11639.16 ± 18922.66 | 0.28 ± 1.11 | 1.89 ± 4.82 | 0.3 ± 0.23 | 0.74 ± 0.71 | 0.13 ± 0.13 |
| **Depression** | 7766.64 ± 13297.92 | 0.43 ± 5.46 | 2.62 ± 8.65 | 0.35 ± 0.23 | 0.72 ± 0.58 | 0.16 ± 0.14 |
| **MDD** | 11163.03 ± 20733.11 | 0.61 ± 2.22 | 3.62 ± 11.1 | 0.33 ± 0.21 | 0.62 ± 0.44 | 0.14 ± 0.14 |
| **OCD** | 6597.84 ± 10579.73 | 0.41 ± 2.06 | 3.4 ± 8.32 | 0.42 ± 0.27 | 0.72 ± 0.49 | 0.17 ± 0.14 |
| **PPD** | 6811.65 ± 12112.21 | 0.29 ± 0.8 | 2.28 ± 3.71 | 0.34 ± 0.22 | 0.74 ± 0.68 | 0.15 ± 0.15 |
| **PTSD** | 7410.33 ± 14476.89 | 0.32 ± 1.16 | 2.53 ± 5.37 | 0.36 ± 0.25 | 0.9 ± 1.07 | 0.16 ± 0.16 |
| **Control** | 9845.13 ± 36824.34 | 0.37 ± 1.91 | 1.89 ± 5.87 | 0.25 ± 0.22 | 0.82 ± 1.03 | 0.16 ± 0.22 |

Table 3: Collected posts aggregate statistics

have lower levels of engagement on the social platform as some previous research works suggest.This shows uniform and unbiased nature of our dataset.

Twitter-STMHD thus qualifies for various feature extraction experiments, and the timeline considered ensures that trends can be captured and analyzed.

## Exploratory Data Analysis

An extensive analysis of the data was carried out in order to determine patterns and differences amongst various sets of users. In order to get an idea of the linguistic style and word usage, the LIWC tool was employed on our dataset. LIWC is an application that consists of a dictionary and counts words in psychologically meaningful categories (Tausczik and Pennebaker 2010). Thus each word in the target text is searched in the dictionary and if there is a match, the count under the appropriate category is incremented. It helps in capturing emotional, cognitive, and structural components present in individuals' writings. (Pennebaker et al. 2015)

Figure 2 depicts the density of positive and negative sentiment tweets. It serves the basic intuition regarding the prevalence of more negative sentiment tweets in case of disorder class users while vice-versa in the case of control users (Rosa et al. 2016). Making it quite evident that users suffering from some form of mental health disorder will have a higher negative connotation in their tweets which could serve as a determining factor in mental health diagnosis.

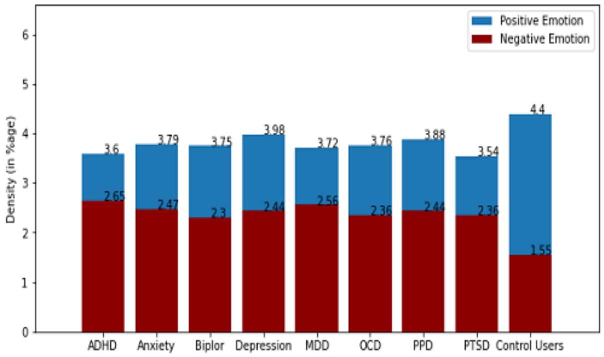

Figure 2: LIWC positive emotion category comparison

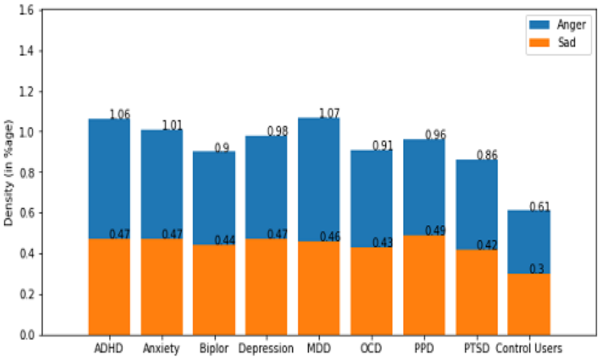

Figure 3: LIWC anger and sad emotion comparison

Moreover, a person suffering from a mental health disorder might have more phases of anger and sadness in comparison to the other. Figure 3 shows that users suffering from mental health disorders had more usage of words depicting anger and sadness in comparison to control class users. The use of personal pronouns is more frequent in people suffering from trauma. In a study, the use of 'I' was prevalent in essays written by depressed users in contrast to those written by non-depressed ones. (Rude, Gortner, and Pennebaker 2004). Figure 4 confirms that control users have less tendency to use personal pronouns in tweets. The LIWC lexicon for 'Home' has words associated with household concerns, the higher count of home-related tweets in Figure 5 for users suffering from a mental health disorder indicates prevailing self-centeredness and homesickness.

Leisure activities such as exercising and other recreational activities play a prominent role in mood enhancement and also in the treatment of depression (Anderson and Brice 2011) (Patten et al. 2013) (Goodman, Geiger, and Wolf 2016). As shown in the Figure 5, our dataset confirms this statement, it shows that control users mentioned leisure activity-related terms in their tweets more frequently as compared to users suffering from a mental health disorder.

As shown in the Figure 6, we tried to observe temporal user posting activity, the frequency of posting was more in night hours in case of users suffering from mental health disorder which aligned with the previous observation in Shen

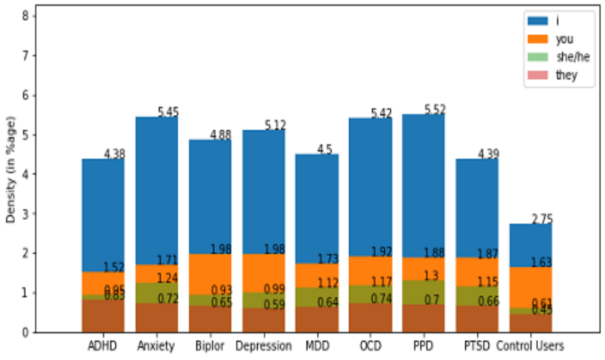

Figure 4: LIWC personal pronouns category comparison

et al.'s work. Insomnia could serve as a risk factor in instigating mental health disorders such as depression (Riemann and Voderholzer 2003) and thus the posting activity of the users could be used as an indicative sign in early detection of mental health disorders.

Further, Figure 7 depicts a sudden surge in posting frequency post March 2020, the time COVID-19 was declared as a pandemic and social-distancing was imposed worldwide. Henceforth our dataset was capable enough to capture people's extensive tweeting activity during this period and it could help in establishing a pattern between COVID-19 and depleting mental health.

## Fairness

The collected data consists of publicly available information from a widely used public social network platform, Twitter. We took care of the FAIR principles while constructing the dataset (Hagstrom 2014). The dataset is **findable**, as it is has been hosted on a data publishing service, Zenodo which assigned the data set a DOI $10.5281/zenodo.6409736$ to aid findability. The dataset can be requested for use through a Data Usage Agreement (DUA) protecting the concerned users' privacy. Our dataset contains nine broad classes of users, one for each of the eight mental health disorders considered and one for control users. For every user, data is provided in json files, one for user-profile information and one for timeline tweets information. An additional json containing the anchor tweets of the eight disorder class users has been given for visual and contextual aid.These practices ensure an **accessible** dataset. The JSON file format makes the data **interoperable**, given that the majority of the current programming languages and softwares have tools and libraries to process files in this format. This dataset is **reusable** as a README file is included with the dataset that explains it in detail. The data we collected was stored in a central server with restricted access and firewall protection. All experiments shown in this paper were conducted on this dataset.

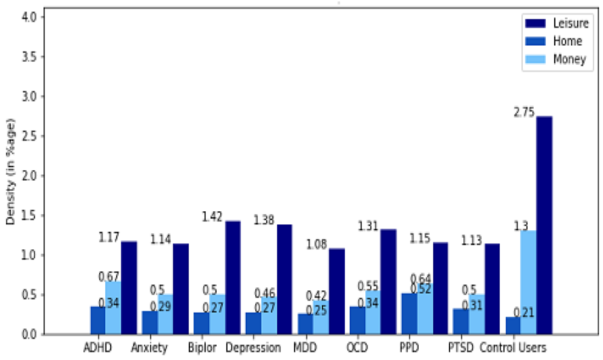

Figure 5: LIWC Leisure, Money, Home category comparison

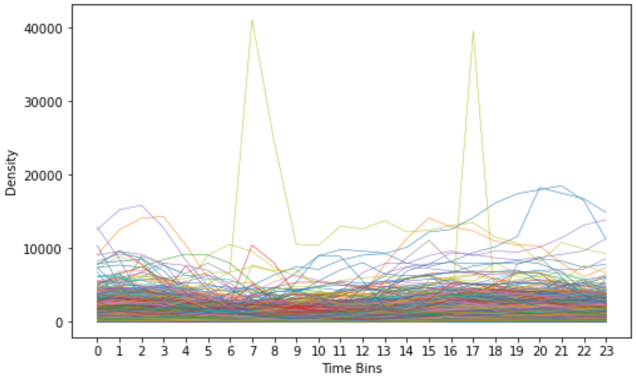

Figure 6: Depression diagnosed user hourly posting activity

## Conclusion and Future Scope

We presented Twitter-STMHD, a large temporal dataset of Twitter users with various mental health conditions and matched control users. Our dataset was collected and constructed following ethical protocols and keeping up with the data quality standards. To our knowledge, Twitter-STMHD is the largest dataset of users suffering from mental health disorders, capturing their tweeting activity over the 2017-2021 timeline. We tried to analyze and extracted various patterns to establish the relevance of our dataset in diagnosis of mental health disorders.

Our dataset captures a user's tweeting activities for the period leading up to their mental health disorder diagnosis. Hence, our next goal is to come up with models which can utilize the temporal and emotional context of data and make accurate predictions regarding the mental health condition of a user. We aim to leverage the COVID-19 data to establish a correlation between the pandemic and the repercussions it had on people's mental health.

## Ethical Statement

It is important to mention that our research and analysis relied on publicly available data that is accessible and collected without interacting with the users who were suffering from mental health conditions. Analysis of publicly available data that could suggest current or prospective mental health disorders poses privacy concerns as well as broader

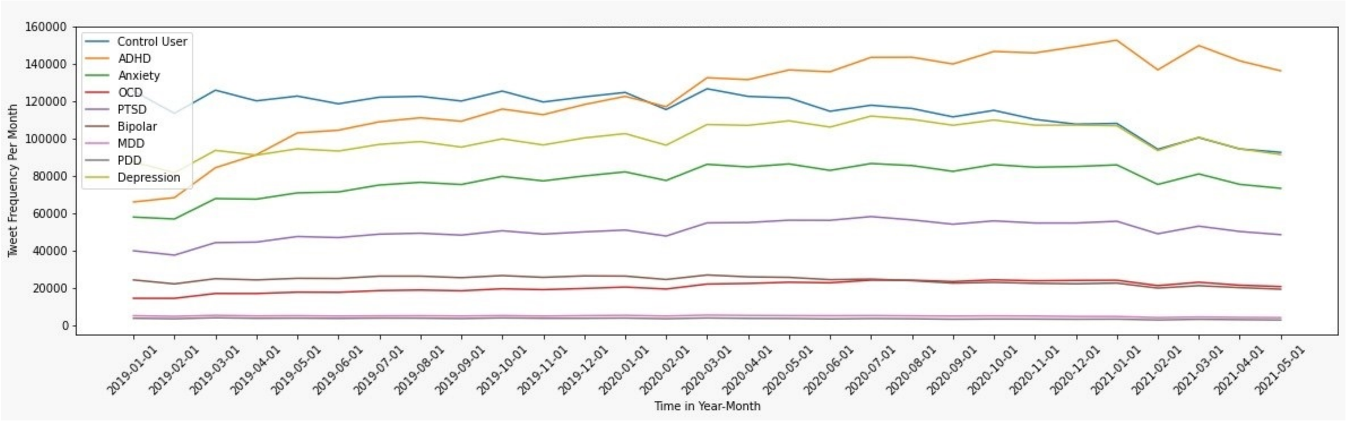

Figure 7: Tweet Trend

ethical questions about undertaking research in this area. To decrease traceability, we took great care in how the data and analyses were presented in the study for each disorder, for example, by omitting any personally identifying information such as tagged users or web-links in the data we provide. Several discussions around ethical considerations for using Twitter dataset concluded that it can ethically be used for research as it is one of the expected use cases of data that Twitter users agree to in terms and conditions. [5]

## Acknowledgements

We would like to thank Dr. Rajiv Ratn Shah and Dr. Ponnurangam Kumaraguru for their invaluable inputs that shaped the project into its present form. Dr. Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT Delhi. Our work is further supported by Midas-IIITD and Dr. PK's Lab, Precog without which the data collection and processing would have been an impossible task. We would like to especially thank Dr. Shefali Gupta, Clinical Psychologist and Assistant Professor at Amity University, Gwalior, for helping us understand the sensitivity of the data we were dealing with and validating our annotations which ensured the credibility of our dataset.

## References

A.; Frances, A.; and APA. 1994. *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV.* Diagnostic and Statistical Manual of Mental Disorders Series. American Psychiatric Association. ISBN 9780890420621.

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, 183–194.

Anderson, R. J.; and Brice, S. 2011. The mood-enhancing benefits of exercise: Memory biases augment the effect. *Psychology of Sport and Exercise*, 12(2): 79–82.

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.

Clement, J. 2018. Countries with most Twitter users 2018. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. Accessed:2022-01-15.

Cohan, A.; Desmet, B.; Yates, A.; Soldaini, L.; MacAvaney, S.; and Goharian, N. 2018. SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1485–1497. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014a. 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. Baltimore, Maryland, USA: Association for Computational Linguistics.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014b. 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.; and Hollingshead, K. 2015a. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, 1–10.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015b. 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.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015c. 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. Denver, Colorado: Association for Computational Linguistics.

[5] https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf

Coppersmith, G.; Ngo, K.; Leary, R.; and Wood, A. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, 106–117.

DataReportal. 2021. Global Social Media Stats. https://datareportal.com/social-media-users. Accessed: 2022-01-15.

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 *Seventh international AAAI conference on weblogs and social media*.

GDB. 2021. global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis from the global burden of disease study. https://www.healthdata.org/research-article/global-regional-and-national-burden-12-mental-disorders-204-countries-and. Accessed: 2022-01-15.

Goodman, W. K.; Geiger, A. M.; and Wolf, J. M. 2016. Differential links between leisure activities and depressive symptoms in unemployed individuals. *Journal of clinical psychology*, 72(1): 70–78.

Hagstrom, S. 2014. The FAIR Data Principles. https://www.force11.org/group/fairgroup/fairprinciples. Accessed: 2022-01-15.

IHME. 2021. New Global Burden of Disease analyses. https://www.healthdata.org/acting-data/new-ihme-analyses-show-depression-and-anxiety-among-top-causes-health-burden-worldwide. Accessed: 2022-01-15.

James, S. L.; Abate, D.; Abate, K. H.; Abay, S. M.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159): 1789–1858.

Kendler, K. S. 2013. A history of the DSM-5 Scientific Review Committee. *Psychological Medicine*, 43(9): 1793–1800.

Kumar, M.; Dredze, M.; Coppersmith, G.; and De Choudhury, M. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*, 85–94.

Lin, Y. 2020. 10 twitter statistics every marketer should know in 2020 [infographic]$_2$020.. Accessed: 2022-01-15.

MacAvaney, S.; Desmet, B.; Cohan, A.; Soldaini, L.; Yates, A.; Zirikly, A.; and Goharian, N. 2018. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 168–173. New Orleans, LA: Association for Computational Linguistics.

Organization, W. H. 2019. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed: 2022-01-15.

Park, M.; Cha, C.; and Cha, M. . 2012. Depressive moods of users portrayed in Twitter. In *Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, SIGKDD 2012*, 1–8.

Park, S.; Lee, S. W.; Kwak, J.; Cha, M.; and Jeong, B. 2013. Activities on Facebook reveal the depressive state of users. *Journal of medical Internet research*, 15(10): e217.

Patten, S.; Williams, J.; Lavorato, D.; and Bulloch, A. 2013. Recreational Physical Activity Ameliorates Some of the Negative Impact of Major Depression on Health-Related Quality of Life. *Frontiers in Psychiatry*, 4.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of LIWC2015. Technical report.

Pfefferbaum, B.; and North, C. S. 2020. Mental health and the Covid-19 pandemic. *New England Journal of Medicine*, 383(6): 510–512.

Riemann, D.; and Voderholzer, U. 2003. Primary insomnia: a risk factor to develop depression? *Journal of affective disorders*, 76(1-3): 255–259.

Rosa, R. L.; Rodríguez, D. Z.; Schwartz, G. M.; de Campos Ribeiro, I.; and Bressan, G. 2016. Monitoring system for potential users with depression using sentiment analysis. In *2016 IEEE International Conference on Consumer Electronics (ICCE)*, 381–382.

Rude, S.; Gortner, E.-M.; and Pennebaker, J. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8): 1121–1133.

Saloni Dattani, H. R.; and Roser, M. 2021. Mental Health. *Our World in Data*. Https://ourworldindata.org/mental-health.

Santomauro, D. F.; Herrera, A. M. M.; Shadid, J.; Zheng, P.; Ashbaugh, C.; Pigott, D. M.; Abbafati, C.; Adolph, C.; Amlag, J. O.; Aravkin, A. Y.; et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, 398(10312): 1700–1712.

Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Stillwell, D.; Kosinski, M.; Ungar, L.; and Schwartz, H. A. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1146–1151. Doha, Qatar: Association for Computational Linguistics.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; and Zhu, W. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In *IJCAI*, 3838–3844.

Singh, S.; Dixit, A.; and Joshi, G. 2020. "Is compulsive social media use amid COVID-19 pandemic addictive behavior or coping mechanism? *Asian journal of psychiatry*, 54: 102290.

Statista. 2021. Global Twitter user age distribution 2019 — Statista. https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/. Accessed: 2022-01-15.

Stennis, C. 2018. Defining what makes Twitter's audience unique. https://blog.twitter.com/en_us/topics/insights/2018/defining-what-makes-twitters-audience-unique. Accessed: 2022-01-15.

Tausczik, Y. R.; and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1): 24–54.

Wang, X.; Zhang, C.; Ji, Y.; Sun, L.; Wu, L.; and Bao, Z. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 201–213. Springer.

WHO. 2019. The New Normal. https://www.who.int/westernpacific/emergencies/covid-19/information/covid-19-new-normal. Accessed: 2022-01-15.