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

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

[1]Indraprastha Institute of Information Technology, Delhi, India
[2]Guru Gobind Singh Indraprastha University, Delhi, India
[3] Indian Institute of Information Technology, Dharwad, India
[4] International Institute of Information Technology, Hyderabad, India

Abstract

Social Media is equipped with the ability to track and quantify user behaviour, establishing it as an appropriate resource for mental health studies. However, previous efforts in the area have been limited by the lack of data or contextually relevant information. There is a need for large-scale, well-labelled mental health datasets with fast reproducible methods to facilitate a heuristically growing dataset. 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, contributing to the rest of the dataset. This dataset records user-profiles and all tweets from relevant periods like onset and progression of disorder by leveraging upon temporal data. This is the only largest dataset that tries to capture the tweeting activity of users suffering from mental health disorders during the COVID-19 period.

Introduction

Depression. Anxiety. Bipolar disorder. Obsessive Behaviours. Trauma. These are a few commonly known mental health illnesses or disorders, which alter the sufferer's emotions, mood, thought behaviour, altering their entire lives, yet lacking clear physical evidence indicating 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 (Collaborators et al. 2022).

World Health Organization states that in its worst form, mental health disorders can lead to suicide ideation, the second leading cause of death among 15-29-year-olds over 2007-2017 (of Mental Health et al. 2005), emphasising the importance of early diagnosis.

Clinically, professional psychologists diagnose mental health disorders in face-to-face interviews, using DSM-V (published in 2013) (The Diagnostic and Statistical Manual of Mental Disorders) as the reference.

The DSM-IV (predecessor of DSM-V [till 2013, published in 1994] (Kendler 2013) describing mental Health disorders notes, "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: 1. unawareness and lack of resources (prevalent in mid to low income countries)[/cite] ii) long-standing social stigma making it a taboo, and iii) imperfect recall of mood and behavioural changes over the observation period (mood and behavioural reports from patients lead to diagnosis in interviews 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.

On the other hand, over the past decade, people have taken to social media platforms like Facebook, Twitter, 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.

The COVID-19 outbreak forced people inside their homes, cutting them from physical worlds, escalating the age of social media, making them resort solely to the platforms for socialising, emoting and interacting with people. A study Singh, Dixit, and Joshi (2020) analysing 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, people's Mental Health under the "new normal" declined drastically (Singh, Dixit, and Joshi 2020) (Pfefferbaum and North 2020). Santomauro et al.

*These authors contributed equally.

(2021) in its paper calls 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. The DSM-V contains descriptions, symptoms, and other criteria for diagnosing mental disorders, along with an approximate period of symptom and disease prevalence. (like two years for Bipolar Disorder and six months for Anxiety Disorders). We maintained the above as the skeleton aim for our study.

The realisation of social media as an important mental health resource gained popularity in the past decade, leading to novel analytical studies and datasets customised 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 microblogging 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. (2015b) 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 behaviours in the period that marked the onset of the disease.

Temporal reference for tweets arms 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, disease prevalence 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. The study 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 a given tweet, all for a period of 3-4 before the onset (date of diagnosis tweet) thus making depression about a person instead of a tweet.

We extend from the previous efforts for addressing the need for labelled 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 "TSTMUD" for many users picked via self-disclosure tweets, grouped into eight mental-health disorder groups and a corresponding class for control users. The Dataset, instead of being a collection of textual posts, is a collection of user profiles. We identified two broad periods of relevant data for any user, before (onset of disease) and after (progression and prevalence of disease), 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 third temporally relevant period, the post-COVID-19 period, to facilitate studies on understanding the stark negative effect of the pandemic on mental health.

Our contribution to the mental-health research space is as follows: We provide a large dataset of users grouped into 8-different mental health disorder classes (identified by self-disclosure tweets) along with a control-users class. A high precision 60Consequently, we have built a lexicon of specialised patterns for accurately identifying self-reported diagnosis reports eliminating the need for manual annotation hence paving the way for a heuristic-growth of Twitter-STHMD. We created 40Realising the need for a temporal context for user activities, we identified three main periods for collecting user-timeline and profile data. (onset of the disorder, progression and disease prevalence period (after the tweet), the pandemic) A user-level dataset instead of a tweet-level corpus dataset, thus putting the complete user behaviour into context similar to building a case history in a clinical diagnosis interview. We have assembled the information into possible feature groups, quantifying the data and the trends in data confirming findings in previously conducted research, validating our methods. Lastly, We provide the first ever large-scale mental health dataset from the pandemic period, noted for deterring mental health at large. Capturing two kinds of users , those with new diagnosis reports during the pandemic and those with diagnosis reports from before it (2017-19), quantifying the implications of the pandemic on their mental health. Our dataset effectively captures how a global-level, real-world change affected both control 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, downvoting, and commenting. Mental health conditions are primarily concerned with behavioural and mood changes and social media became a medium that captured a users' behavioural aspects in a quantifiable way. The connection between the two is indisputable. Coppersmith, Dredze, and Harman (2014a) 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 behavioural 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 (like: I have been diagnosed with depression) on Twitter (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. 2013) 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 periods (post-childbirth). Their classification framework accuracy jumped from 71% to 80-83% by simply leveraging behavioural 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. (2015a) constructed a large dataset covering 11 unique disorders; however, the dataset looked like 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 analysed for linguistics patterns. The practice, by default, assumes the possibility of a possible 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 (as considered only 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. 2018) 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. 2017) used self-disclosed diagnosis tweets to build a user-level dataset of about 3600 depressed category users. The dataset contains two types of information for every user: profile statistics and timeline tweets (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 compared for negative-class and depressed-class users. Most notably, the dataset included timestamp-date for all user activity to indicate sleeping patterns, using which as a feature group, the study devised a gold-standard classifier framework.

The user-level granularity of the dataset resolves the issue of low character-cutoff limits on Twitter, as a large number of tweets can be appended sequentially owing to tweet timestamp data with or without temporal weights.

This work identifies the limitations of previous datasets, specifically improving upon [/cite][Shien et al.] for user identification and data collected for any given user. Twitter-STMHD is a large-scale dataset of eight mental health disorders (falling under various categories in DSM-V) classes and one control-user class. This work identifies high-precision patterns to identify valid self-reported diagnosis disclosure tweets (anchor tweets) and eliminate the need for hand annotation. The dataset contains two types of information for every user: user profile data (user.json in data) and timeline-tweets data(tweets.json) for a custom period for every user. The start date is taken two years before the anchor tweet date and the end date, two years after). The dates are adjusted for every user using their anchor tweet timestamp data. The two years before the diagnosis data quantifies the onset phase of a given disorder, and the two years after the diagnosis data quantifies the probable disease progression and prevalence phase. This work aims to cater to the need for a well-labelled, temporally-relevant large-scale mental health dataset to aid research in the domain.

Data

This section explains the construction and features of Twitter-STMHD. The dataset contains eight mental-health disorder classes and a corresponding class of control users unlikely to have any of the disorders discussed in this study. The disorders belong to broader categories of mental health disorders defined in DSM-V. For each user, the dataset provides two types of information, user-profile data and timeline-tweets data, containing all tweets from the user-specific data collection period. The DSM-V constitutes a nomenclature of mental-health disorders. It lists several behavioural impacts of disorders and mentions probable windows of symptom and disease prevalence. To define a custom period for data collection for a given user, we take the maximum of the respective periods mentioned in the DSM

for our considered disorders and disorder categories. The 8 disorders, their common names, and prevalence period are:

1. *Depression* (2 years for adults).
2. *Anxiety* (1 year).
3. *Obsessive-Compulsive Disorder* (OCD).
4. *Bipolar Disorder* (2 years).
5. *Post-Traumatic Stress Disorder* (PTSD), can develop within 1 month of traumatic event.
6. *Major Depressive Disorder* (MDD) (2 years).
7. *PostPartum Depression (PPD)* (1 month prenatal, 1 month post-natal).
8. *Attention-Deficit Hyperactivity Disorder* (ADHD) (2 years) under neurodevelopmental disorders.

Our dataset also includes users among multiple mental health conditions, providing an amazing resource

Dataset Construction

The STMHD dataset was created by selecting users with self-reported diagnosis disclosure posts on Twitter. We will refer to these tweets as anchor tweets for the purpose they serve and the ease of reference. The study uses Twitter for collection because of readily usable APIs that give flexibility with the period for which the dataset needs to be collected, allowing us to customise the period limits for each user. Once identified to belong to a particle disorder class, we extract the user's profile information and the timestamp of the anchor-tweet anchor tweet to define user-specific relevant data collection periods. We then collect all tweets and activity data for the determined period. This approach is based on the kind of information given for the user in the (Shen et al. 2017) paper.

COVID-19 Data COVID-19 was a significant event that affected people worldwide. The period also deterred mental health on a large scale. Simultaneously, social media recorded an exponential increase in activity during the pandemic. The realisation for the need to understand the deterring mental health conditions and the dispense of information at a large rate owing to big traffic of online users prompted us to collect data the peak COVID period, for 2 kinds of user-groups, first, users with anchor tweets from before the pandemic, January 2017 to March 2020. The pandemic user activity for all of these users was collected in addition to their data collection windows. Second, the users with anchor tweets dated post the announcement of the pandemic, March 2020 to May 2021. Effectively, we collected anchor tweets from January 2017 to May 2021 and for the users s identified, collected timeline data for their data collection windows as well as the through the first year of the pandemic (March-2020 to May-2021).In Twitter-STMHD we have identified 25,860 unique users belonging to atleast one of the eight disorder categories and a control user class with more than 8000 users, least likely to have any of the eight disorders.

Anchor Tweets Identification Anchor Tweets are tweets where a user claims to have been diagnosed with one of the eight mental health conditions. These anchor tweets were identified using a combination of manual annotation process and high precision diagnosis patterns; the contribution from the two methods has been listed in Table 3.

Anchor-Tweet Collection Period The anchor tweets were collected between January 2017 and May 2020. This was done in order to capture users in both *pre-COVID-19 period* and *post-COVID-19 period*. The disorder wise distribution of the anchor tweets is listed in Table 3.

User-specific data collection window 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, post the diagnosis. To identify this window, we took the maximum of all periods of observation suggested in the DSM-5 for the 8 disorders in question and the categories of disorders they belong to. Consequently, 4-year window was chosen with 2year of data before and after the anchor(self-reported diagnosis)tweet.

Building High-precision Patterns for Identifying Anchor Tweets While hand-annotation is a time-expensive procedure, it accounts for the most precise and reliable datasets. 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 data corpuses. This explains the need for high-precision patterns for identifying anchor tweets containing the lexicon for identifying valid anchor tweets and as well as clearly invalid ones for maximum precision approach. To identify the anchor tweets for the eight mental health disorders, with high precision , we follow a two-step approach with the ultimate aim of eliminating the need for hand annotations and hence speed-up the process of data collection, paving the way for the heuristical growth of well labelled mental health resources. Earlier approaches relying on self-disclosure post keywords like 'got diagnosed' have proven the method useful for the base selection of users. But this approach is not robust and leads to several false positives like 'My mother got diagnosed with ADHD, 'he got diagnosed with depression etc. First, we manually annotate the base corpus containing both positive and negative instances of a valid anchor tweet. We start by collecting tweets that match the loose pattern "got diagnosed with ¡disorder¿". A typical anchor tweet has two parts: a self-disclosure of diagnosis and the disorder's name. We prepared a lexicon for disorder names using common synonyms, their names in DSM-V, mis-spellings, short forms and full forms and used this lexicon with the loose regex to capture maximum possible probable anchor tweets. The number of collected tweets per user is present in table 1, this set of collected tweets formed our preliminary candidate set of anchor tweets. This candidate set was then divided into two equal parts for each disorder, for hand-annotation and pattern-matching. Data was hand-annotated by 5 contributors. We hand annotated and recognised 26,117 tweets as valid anchor tweets. These tweets recorded 25,439

unique users owing to contributing to 60% of the users in the dataset. To validate our annotations, a sample of 500 tweets containing both positive and negative annotations was sampled by a clinical psychologist , they disagreed with 4 out of 500 annotations. Thus making 60% of our dataset 99.2 % precise. In the annotation process, a lexicon of the observed string patterns in the self-diagnosis tweets of the annotation set was developed for each disorder. To check the performance of the created lexicon, we evaluated its performanve on the hand-annotated dataset, we were able to identify the positive users with a 94% precision considering hand-annotations to be correct. This lexicon of high precision string patterns was to then identify anchor tweets, contributing to 40% of the users in dataset. The self-diagnosed tweets from the filtered and the lexicon set are used to create the final dataset of anchor tweets. Table 1 records the final counts of disorder wise anchor tweets from both hand-annotation adn pattern-matching

Disorder Dataset Users Diagnosed users are users who posted an anchor tweet in the period between January 2017 and May 2020 are put into their respective disorder categories. From the identified anchor tweets the Twitter IDs of the users were extracted to make our preliminary candidate set of self-reported diagnosis users for each mental health condition. For each of the users in the user set, all user activity was collected between periods T1 and T2, where T1 and T2 is calculated as follows:

$$\Delta = max(T_{depression}, T_{anxiety}, T_{ocd}, T_{bpd}, T_{ptsd}, T_{mdd}, T_{ppd}, T_{adhd}) = 2\ years$$

$$T1 = T_{anchor} - \Delta \quad (1)$$

$$T2 = min(T_{anchor} + \Delta, May2021) \quad (2)$$

We choose a $\Delta = 2\ years$ before and after the self diagnosis tweet to capture the tweets generated by the user. The period of two years is in line with the DSM 5's symptom observation period recommendation. We stopped our data collection in May 2021, since that period marked the end of major waves of the pandemic around the world, and the period was also close to the time of wrapping up our collection of probable anchor tweets using the loose query statement: "diagnosed with [disorder name]". After the collection of the candidate self-reported diagnosis mental health disorder users, we filter the set to come up with our final self-reported diagnosis user set.

Weeding out undesirable user accounts User Profiles were weeded out based on 2 factors:

1. *Min. Tweet Count*: Any user with less than 50 tweets for the period of 4 years around the anchor tweet timestamp was removed. This is done to ensure that enough tweets are collected to perform any analysis that wants to capture the temporal differences between the tweets and wants to utilise the context of the tweets.
2. *Max. Follower Count*: Any user with a follower count greater than 5000 are also removed. This is done to ensure we do not have any accounts pertaining to famous personalities and influencers that could potentially use Twitter as a brand advertisement tool, and also mental health and other wellness organisations using Twitter as a medium for wellness awareness, indulging in health discourse primarily.

The weeding process ensures our dataset captures the general users on the platform who engage in natural and organic conversations. The remaining users are included in our final dataset. Numbers have been noted in 4.

Control Userset Creation We excluded any personal bias in deciding the date range for the collection of control users. To ensure temporal consistency, we collected tweets, randomly sampled in the same range as the positive class (between January 2017 and May 2020). For each tweet, we identified the user and collected its data starting from two years prior to the posting time of the tweet till May 2021. From the users for which we collected the data, we weeded out users who had any evidence of mental health discourse in their tweets. To carry out this task, we made a set of lexicons pertinent to mental health and topics around it. We checked the presence of lexicons in control user tweets for them to be least likely to belong to our eight disorder sets. Control users, too, were weeded out based on min. Tweet count (50) and max. Follower count (5000) to only capture and keep desirable, high-quality users in the dataset.The control user class contains 8199 distinguished users.

User data collected For each tweet, we collected several essential attributes while ensuring that the user's identity is not compromised and contains no individual identifier. These attributes include:

1. *text*: tweet content.
2. *conversation_id*: unique identifier of the conversation of which the tweet is a part of.
3. *tweet_id*: the unique identifier of the tweet.
4. *language*: the language of the tweet.
5. *Likes_count*: number of likes received by the tweet.
6. *quote_count*: number of times the tweet has been quoted.
7. *Reply_count* : number of replies the tweet has received.
8. *Retweet_count*: number of times the tweet has been retweeted.
9. *source_name*: the source from which the tweet was made.
10. *timestamp_tweet*: the time when the tweet was posted by a user.
11. *mentionedUsers*: the list of user ids which have been mentioned in the tweet.
12. *mentionedUsers*: the list of user ids which have been mentioned in the tweet.
13. media which specifies all the images and videos in the tweet.

For each user, we collected the following attributes:

1. *creation_timestamp*: the timestamp for user creation.
2. *description* : that the user has added to their account.
3. *favorites_count* : number of likes given by the user.

4. *friends_count* : number of friends of the user's account.
5. *follower_count*: number of followers of the user's account.
6. *banner_link* : URL to a downloadable link of banner picture.
7. *display_image_link* : URL to a downloadable link of display image.
8. *status_count*: number of tweets posted by the user's account.
9. *verfiied_check*: noting if the user's account is verified or not.

Tagging Mental Health Discourse in Dataset Additionally, we added a disorder_flag which specifies if the text attribute of the tweet contains words pertaining to mental health and related topics using the same lexicon build to remove users from the control-userset.The date was not removed [/cite][shmd], but flagged for the scope of study of mental health discourse in users.

Scope for multimodality In an effort to give this dataset the scope of multimodality, each tweet from timeline json in the dataset has a "media" category (as shown above), that lists the kind of media (image, gif, video) attached to a tweet and a downloadable link for the same. All profile mappable or cross-platform identity mapping links were removed and replaced with downloadable links to the image or the videa. 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 by the virtue of ethical considerations.

statistics The collected user features and posts features are used to calculate aggregate statistics. Statistics for users can be found in table 2, and statistics for Posts can be found in table 1.

Data Quality

Twitter is a social and microblogging platform which boasts around 192 million daily active users (Twitter, 2021). As of April 2021, Twitter's global audience was composed of 38.5 percent of users aged between 25 and 34 years old, 21 percent users aged between 35 and 49, 24 percent users below 24 years and users aged 50 or above accounted for around 17 percent [1]. Twitter roughly has 34 percent female and 66 percent male [2]. It has the most number of users from the US around 77 million followed by Japan, India, Brazil, the UK and several other countries [3]. All of these points make twitter a demographically rich medium to collect our data and moreover the readily available API of twitter which puts 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. [4] 78% of US 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 user for advice when they want to learn more about a topic [5]. Henceforth no doubt we can say twitter has quality users with active engagement. User relationships and activity is a determining factor in establishing social media dataset quality [6]. Thus in order to get an overview of user tweeting activity and relationships for our dataset, we found out the average number of tweets done by any user per year in a particular disorder class and an average number of followers and friends she/he has.

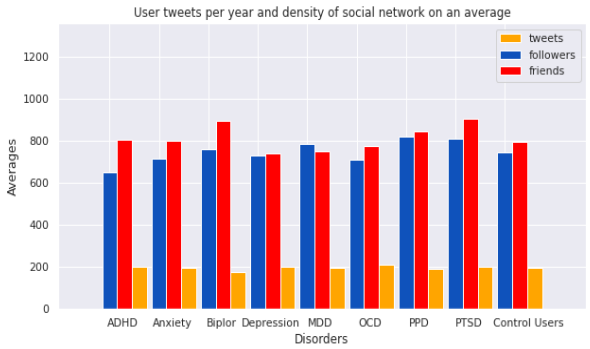

Figure 1: Users Relationship and Activity

Figure 1. depicts that the control users had a similar distribution to other disorders which in turn reflects the quality of users collected, as on average it doesn't have drastically low values for any disorder class which could lead to the wrong hypothesis that there is lower level of engagement on the social platform for user suffering from mental health disorder as some previous research work try to point which might not entirely be true, moreover it shows uniform and unbiased nature of our dataset. In addition, the dataset discussed in this section qualifies for various feature extractions, and the timeline considered ensures that trends can be captured and analyzed.

Exploratory Data Analysis

After collecting tweets of users pertaining to eight major disorders and control users, 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 amongst users belonging to various sets of disorders, the LIWC tool was employed on our dataset. LIWC is an application that consists of a dictionary and counts words in psychologically meaningful categories. [7] Thus each word in the target text is searched in

[1] https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

[2] https://www.oberlo.in/blog/twitter-statistics

[3] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[4] https://www.oberlo.in/blog/twitter-statistics

[5] https://blog.twitter.com/en_us/topics/insights/2018/defining-what-makes-twitters-audience-unique

[6] https://dl.acm.org/doi/pdf/10.1145/1341531.1341557

[7] https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf

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 1: Collected posts aggregate statistics

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

Disorder	Collected Tweets	Final Anchor Tweets	Final Anchor Tweets	
			Hand Annotated	Pattern Annotated
ADHD	43764	8688	5039	3649
Depression	37149	11133	6791	4342
PTSD	30077	5009	3155	1854
Anxiety	26739	9654	5985	3669
OCD	7558	2320	1415	905
PPD	713	596	333	263
MDD	651	510	331	179
Bipolar	5967	5559	3168	2391
Total Counts	152618	43269	26117	17152

Table 3: Represents the vast number of user anchor tweets first identified using a loose regex "diagnosed with disorder" against the number of valid ones (indicating no. of users in each datset) and number of tweets recognised by both hand-annotations and pattern matching.

Disorder	User Counts
Depression	6803
PTSD	3414
Anxiety	4843
OCD	1325
PPD	547
MDD	325
Bipolar	1651
Control Group	8199

Table 4: Final user count for the disorders after weeding out by scanning for minimum >= 50 users and maximum <= 5000 followers.

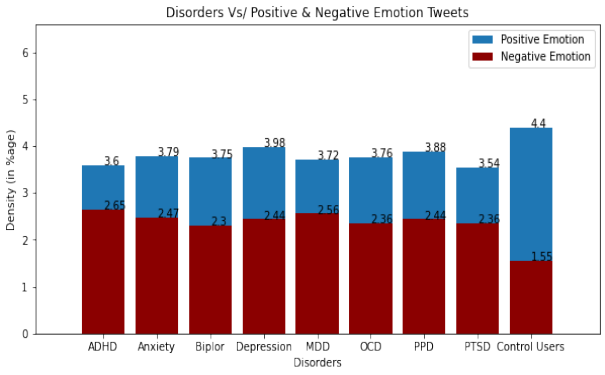

Figure 2: LIWC pos emo category comparison

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. [8]

Figure 2 depicts the density of positive and negative tweets and it serves the basic intuition regarding the prevalence of more negative tweets in case of disorder while vice-versa in the case of control users (Rosa et al. 2016). Thus it's quite evident that a user suffering from some form of mental health disorder will have a significant negative connotation in her/his tweets which could serve as a determining factor in one's health diagnosis.

Moreover, a person suffering from a mental health disorder might have more phases of anger and sadness in comparison to a normal person. Figure 3 and Figure 4 shows that users suffering from mental health disorders had more usage of words depicting anger, sadness and swearing more in their tweets in comparison to control users.

The use of personal pronouns is more frequent in people suffering from trauma and have attention to themselves. In a study use of 'I' was prevalent in essays written by depressed users than in non-depressed ones (Rude, Gortner, and Pennebaker 2004). Figure 5 displays the same that control users had less tendency to use personal pronouns in tweets than

[8]https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

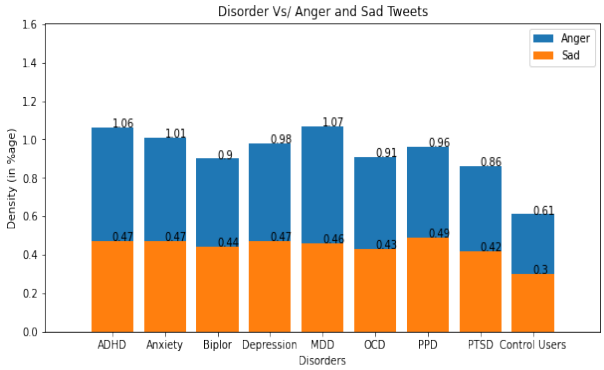

Figure 3: LIWC Anger Sad Emo Comparison

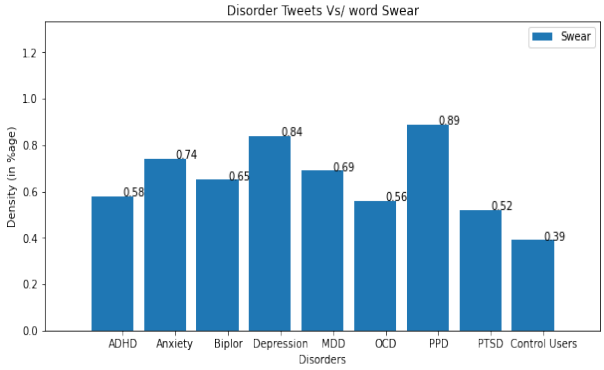

Figure 4: LIWC Swear Word Count Comparison

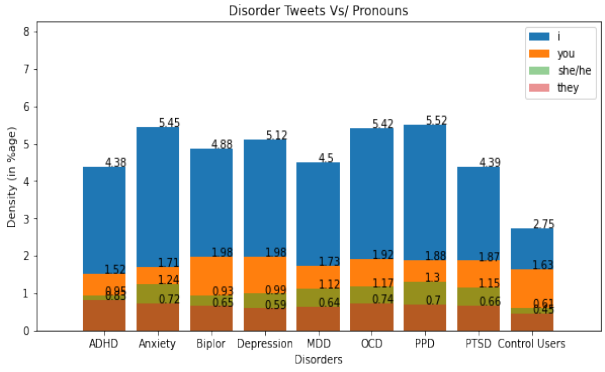

Figure 5: LIWC Personal Pronouns Category Comparison

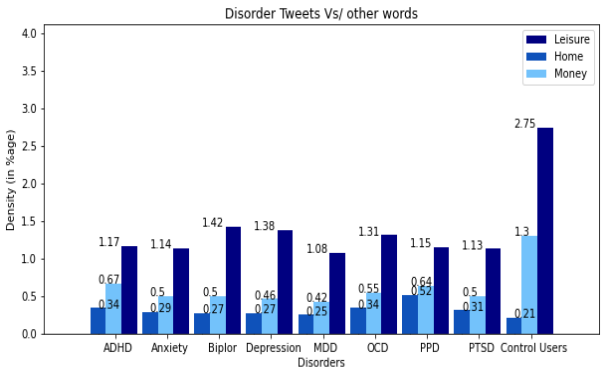

Figure 6: LIWC Leisure, Money, Home category comparison

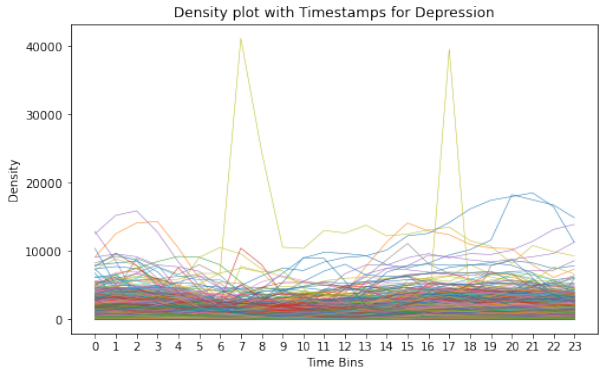

Figure 7: Depression diagnosed User Hourly Posting Activity

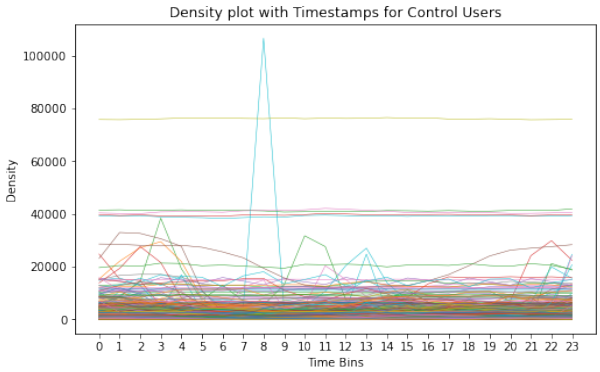

Figure 8: Control User Hourly Posting Activity

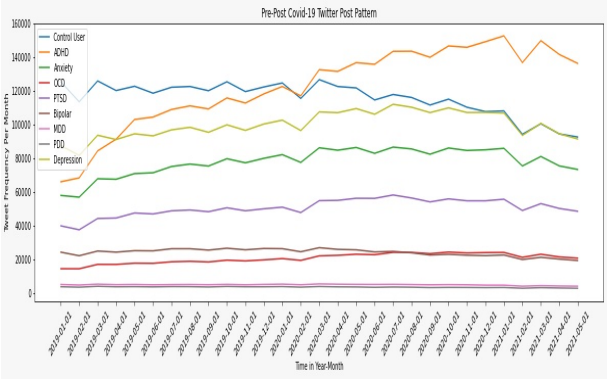

Figure 9: Tweet Trend

in the case of depressed users. Moreover, the use of 'I" was more recurrent in users suffering from a mental health disorder. The LIWC lexicon for Home has words more associated with household concerns, the higher count of home-related tweets in Figure 6 for users suffering from a mental health disorder shows prevailing self-centeredness as discussed above 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 6 Our datasets conform to this statement as it's quite clear that control users mentioned leisure activity-related terms in their tweets more frequently than compared to users suffering from a mental health disorder.

As shown in the Figure 7 and 8, When we tried to observe 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 regarding depressed user more frequent posting activity at night (Shen et al. 2017). (Riemann and Voderholzer 2003) Moreover, insomnia could serve as a risk factor in instigating mental health disorders such as depression and thus the posting activity of the user in our dataset could be used as an indicative sign in early detection of mental health disorder.

Further, Figure 9 depicts a sudden surge in posting frequency from around march 2020 which is the same time COVID-19 was at its peak. 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 one individual mental health status.

Ethics

We should point out that our research and analysis relied on publicly available data that is publicly accessible and collected without interacting with the users, who were dealing with a range of mental conditions, some of which were undiagnosed. Analysis of publicly available data that could suggest current or prospective mental health disorders poses privacy concerns, as well as broader 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, weblinks, personal information, and retweets that we mention. Several discussions around ethical consideration 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. [9]

Fairness

The collected data consists of publicly available information from a widely used public social network platform, Twitter. The FAIR principles are also followed in our dataset[10].

The collected data consists of publicly available information from a widely used public social network platform, Twitter. The FAIR principles are also followed in our dataset[11].

- **Findable**: In particular, the dataset is "findable", as it is has been hosted on a data publishing service, Zenoda which assigned the data set a DOI [10.5281/zenodo.5854911] to aid findability. The dataset can be requested for use, through a Data Usage Agreement (DUA) protecting the concerned users' privacy.
- **Accessible**: Our dataset contains 8 broad classes of users, one for each of the 8 mental health disorders considered and 1 for control user. Fore 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 tweet of the 8 disorder class users has been given visual and contextual aid. These practices ensure an accessible dataset.
- **Interoperable**: The JSON file format also 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.
- **Reusable**: This dataset is also "reusable" as the included README file that explains the dataset 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 performed on this dataset.

Conclusion

We presented STMHD, a large temporal dataset of twitter users with various mental health conditions and matched control users. Our dataset was collected and constructed following ethic protocols and keeping up with the data quality standards. To our knowledge, STMHD is the largest dataset that maintains a database of users suffering from mental health tweeting activity over the 2017-21 timeline. We tried to analyze and extracted various patterns to establish relevance of our dataset in early diagnosis of mental health disorder. For future scope we tend to use various state-of-the-art Language models to predict early signs of mental health disorder in a user. We make our dataset available to the public, and hope that it will facilitate further research in the mental health domain.

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

[10]https://www.force11.org/group/fairgroup/fairprinciples

[11]https://www.force11.org/group/fairgroup/fairprinciples

References

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.

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. *arXiv preprint arXiv:1806.05258*.

Collaborators, G. . M. D.; et al. 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*.

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.

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.

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*.

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.

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.

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.

of Mental Health, W. H. O. D.; Abuse, S.; Organization, W. H.; of Mental Health, W. H. O. D.; Health, S. A. M.; Evidence, W. H. O. M. H.; and Team, R. 2005. *Mental health atlas 2005*. World Health Organization.

Park, M.; Cha, C.; and Cha, M. 2012. Depressive moods of users portrayed in Twitter.

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.

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.

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.