On the State of Social Media Data for Mental Health Research

Keith Harrigian, Carlos Aguirre, Mark Dredze
Johns Hopkins University
Center for Language and Speech Processing

Abstract

Data-driven methods for mental health treatment and surveillance have become a major focus in computational science research in the last decade. However, progress in the domain, in terms of both medical understanding and system performance, remains bounded by the availability of adequate data. Prior systematic reviews have not necessarily made it possible to measure the degree to which data-related challenges have affected research progress. In this paper, we offer an analysis specifically on the state of social media data that exists for conducting mental health research. We do so by introducing an open-source directory of mental health datasets, annotated using a standardized schema to facilitate meta-analysis.[1]

1 Introduction

The last decade has seen exponential growth in computational research devoted to modeling mental health phenomena using non-clinical data (Bucci et al., 2019). Studies analyzing data from the web, such as social media platforms and peer-to-peer messaging services, have been particularly appealing to the research community due to their scale and deep entrenchment within contemporary culture (Perrin, 2015; Fuchs, 2015; Graham et al., 2015). Such studies have already yielded novel insights into population-level mental health (De Choudhury et al., 2013; Amir et al., 2019a) and shown promising avenues for the incorporation of data-driven analyses in the treatment of psychiatric disorders (Eichstaedt et al., 2018).

These research achievements have come despite complexities specific to the mental health space often making it difficult to obtain a sufficient sample size of high-quality data. For instance, behavioral disorders are known to display variable clinical presentations amongst different populations, rendering annotations of ground truth inherently noisy (De Choudhury et al., 2017; Arseniev-Koehler et al., 2018). Scalable methods for capturing an individual's mental health status, such as using regular expressions to identify self-reported diagnoses or grouping individuals based on activity patterns, have provided opportunities to construct datasets aware of this heterogeneity (Coppersmith et al., 2015b; Kumar et al., 2015). Still, they typically rely on stereotypical hypotheses that lack the same clinical validation and robustness as something like a mental health battery (Zhang et al., 2014; Ernala et al., 2019).

Ethical considerations further complicate data acquisition, with the sensitive nature of mental health data requiring tremendous care when constructing, analyzing, and sharing datasets (Benton et al., 2017). Privacy-preserving measures, such as de-identifying individuals and requiring IRB approval to access data, have made it possible to share some data across research groups. However, these mechanisms can be technically cumbersome to implement and are subject to strict governance policies when clinical information is involved due to HIPAA (Price and Cohen, 2019). Moreover, many privacy-preserving practices require that signal relevant to modeling mental health, such as an individual's demographics or their social network, are discarded (Bakken et al., 2004). This missingness has the potential to limit algorithmic fairness, statistical generalizability, and experimental reproducibility (Gorelick, 2006).

Prior systematic reviews of computational research for mental health have noted several of the aforementioned challenges, but have predominantly discussed the technical methods (e.g. model architectures, feature engineering) developed to surmount existing constraints (Guntuku et al., 2017; Wongkoblap et al., 2017). Recent work from Chan-

[1] https://github.com/kharrigian/mental-health-datasets

cellor and De Choudhury (2020), completed concurrently with our own, was the first review to focus specifically on the shortcomings of *data* for mental health research. Our study independently affirms the findings of Chancellor and De Choudhury (2020) using a moderately expanded pool of literature that more acutely focuses on *language* found in social media data. We also offer additional recommendations regarding future dataset curation, supporting our positions using a new open-source directory of mental health datasets.

2 Data

To generate evidence-based recommendations regarding mental health dataset curation, we require knowledge of the extant data landscape. Unlike some computational fields which have a surplus of well-defined and uniformly-adopted benchmark datasets, mental health researchers have thus far relied on a decentralized medley of resources. This fact, spurred in part by the variable presentations of psychiatric conditions and in part by the sensitive nature of mental health data, thus requires us to compile a new database of literature. In this section, we detail our literature search, establish inclusion/exclusion criteria, and define a list of dataset attributes to analyze.

2.1 Dataset Identification

Datasets were sourced using a breadth-focused literature search. After including data sources from the three aforementioned systematic reviews (Guntuku et al., 2017; Wongkoblap et al., 2017; Chancellor and De Choudhury, 2020), we searched for literature that lie primarily at the intersection of natural language processing (NLP) and mental health communities. We sought peer-reviewed studies published between January 2012 and December 2019 in relevant conferences (e.g. NAACL, EMNLP, ACL, COLING), workshops (e.g. CLPsych, LOUHI), and health-focused journals (e.g. JMIR, PNAS, BMJ).

We searched Google Scholar, ArXiv, and PubMed to identify additional candidate articles. We used two search term structures — 1) (mental health | `DISORDER`) + (social | electronic) + media, and 2) (machine learning | prediction | inference | detection) + (mental health | `DISORDER`). '|' indicates a logical or, and `DISORDER` was replaced by one of 13 mental health keywords.[2] Additional literature was identified using snowball sampling from the citations of these papers. To moderately restrict the scope of this work, computational research regarding neurodegenerative disorders (e.g. Dementia, Parkinson's Disease) was ignored.

2.2 Selection Criteria

To enhance parity amongst datasets considered in our meta-analysis, we require datasets found within the literature search to meet three additional criteria. While excluded from subsequent analysis, datasets that do not meet this criteria are maintained with complete annotations in the aforementioned digital directory. In future work, we will expand our scope of analysis to reflect the multi-faceted computational approaches used by the research community to understand mental health.

1. Datasets must contain non-clinical electronic media (e.g. social media, SMS, online forums, search query text)

2. Datasets must contain written language (i.e. text) within each unit of data

3. Datasets must contain a dependent variable that captures or proxies a psychiatric condition as defined by the DSM-5 (APA, 2013).

Our first criteria excludes research that examines electronic health records or digitally-transcribed interviews (Gratch et al., 2014; Holderness et al., 2019). Our second criteria excludes research that, for example, primarily analyzes search query volume or mobile activity traces (Ayers et al., 2013; Renn et al., 2018). It also excludes research based on speech data (Iter et al., 2018). Our third criteria excludes research in which annotations are only loosely associated with their stated mental health condition. For instance, we filter out research that seeks to identify diagnosis dates in self-disclosure statements (MacAvaney et al., 2018), in addition to research that proposes using sentiment as a proxy for mental illness (Davcheva et al., 2019). This last criteria also inherently excludes datasets that lack annotation of mental health status altogether (e.g. data dumps of online mental health support platforms and text-message counseling services) (Loveys et al., 2018; Demasi et al., 2019).

[2] Depression, Suicide, Anxiety, Mood, PTSD, Bipolar, Borderline Personality, ADHD, OCD, Panic, Addiction, Eating, Schizophrenia

2.3 Annotation Schema

We develop a high-level schema to code properties of each dataset. In addition to standard reference information (i.e. Title, Year Published, Authors), we note the following characteristics:

- **Platforms**: Electronic media source (e.g. Twitter, SMS)

- **Tasks**: The mental health disorders included as dependent variables (e.g. depression, suicidal ideation, PTSD)

- **Annotation Method**: Method for defining and annotating mental health variables (e.g. regular expressions, community participation/affiliation, clinical diagnosis)

- **Annotation Level**: Resolution at which ground-truth annotations are made (e.g. individual, document, conversation)

- **Size**: Number of data points at each annotation resolution for each task class

- **Language**: The primary language of text in the dataset

- **Data Availability**: Whether the dataset can be shared and, if so, the mechanism by which it may be accessed (e.g. data usage agreement, reproducible via API, distribution prohibited by collection agreement)

If a characteristic is not clear from the dataset's associated literature, we leave the characteristic blank; missing data points are denoted where applicable. While we simplify these annotations for a standardized analysis — e.g. different psychiatric batteries used to annotate depression in individuals (e.g. PHQ-9, CES-D) are simplified as "Survey (Clinical)" — we maintain specifics in the digital directory.

3 Analysis

Our literature search yielded 139 articles referencing 111 nominally-unique datasets. Application of exclusion criteria left us with 102 datasets. Figure 1 presents the number of articles remaining in our evaluation after these filtering criteria were applied, in addition to filtering criteria related to data availability.

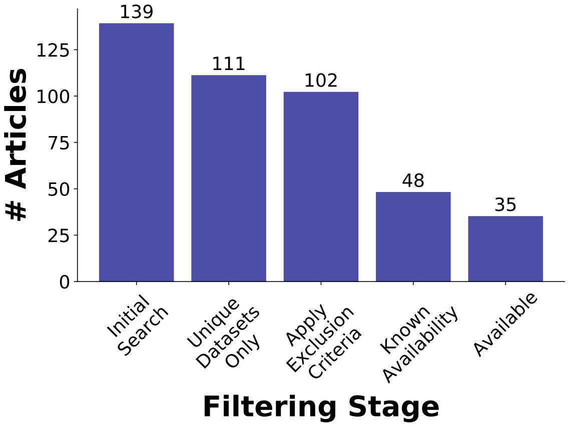

Figure 1: Number of articles (e.g. datasets) remaining after each stage of filtering. We were unable to readily discern the external availability of supporting datasets for over half of the studies.

A majority of the datasets were released after 2012, with an average of 12.75 per year, a minimum of 1 (2012) and a maximum of 23 (2017). The 2015 CLPsych Shared Task (Coppersmith et al., 2015b), Reddit Self-reported Depression Diagnosis (Yates et al., 2017), and "Language of Mental Health" (Gkotsis et al., 2016) datasets were the most reused resources, serving as the basis of 7, 3, and 3 additional publications respectively. Table 1 lists all datasets known to be available for distribution (along with a subset of our annotations). Datasets unavailable for distribution and those with unknown availability may be found in our digital directory with similar annotations.

Platforms. We identified 20 unique electronic media platforms across the 102 datasets. Twitter (47 datasets) and Reddit (22 datasets) were the most widely studied platforms. YouTube, Facebook, and Instagram were relatively underutilized for mental health research — each found less than ten times in our analysis — despite being the three most-widely adopted social media platforms globally (Perrin and Anderson, 2019). We expect our focus on NLP to moderate the presence of YouTube and Instagram based datasets, though not entirely given both platforms offer expansive text fields (i.e. comments, tags) in addition to their primary content of video and images (Chancellor et al., 2016a; Choi et al., 2016). It is more likely that use of these platforms for research, as well as Facebook, is hindered by increasingly stringent privacy policies and ethical concerns (Panger, 2016; Benton et al., 2017).

Tasks. We identified 36 unique mental health related modeling tasks across the 102 datasets.

Reference	Platform(s)	Task(s)	Level	# Inds.	# Docs.	Availability
Coppersmith et al. (2014a)	Twitter	BI, PTSD, SAD, DEP	Ind.	7k	16.7M	DUA
Coppersmith et al. (2014b)	Twitter	PTSD	Ind.	6.3k	-	DUA
Jashinsky et al. (2014)	Twitter	SI	Doc.	594k	733k	API
Lin et al. (2014)	Twitter, Sina Weibo, Tencent Weibo	STR, STRS	Ind.	23.3k	490k	API
Coppersmith et al. (2015a)	Twitter	ANX, EAT, OCD, SCHZ, SAD, BI, PTSD, DEP, ADHD	Ind.	4k	7M	DUA
Coppersmith et al. (2015b)	Twitter	PTSD, DEP	Ind.	1.7k	-	DUA
De Choudhury (2015)	Tumblr	EAT, EATR	Ind.	28k	87k	API
Kumar et al. (2015)	Reddit, Wikipedia	SI	Ind.	66k	19.1k	API
Mowery et al. (2015)	Twitter	DEP	Doc.	-	129	AUTH
Chancellor et al. (2016b)	Tumblr	EATR	Ind.	13.3k	67M	API
Coppersmith et al. (2016)	Twitter	SA	Ind.	250	-	DUA
De Choudhury et al. (2016)	Reddit	PSY, EAT, ANXS, SH, BI, PTSD, RS, DEP, PAN, SI, TRA	Ind.	880	-	API
Gkotsis et al. (2016)	Reddit	ANX, BPD, SCHZ, SH, ALC, BI, OPAD, ASP, SI, AUT, OPUS	Ind.	-	-	API
Lin et al. (2016)	Sina Weibo	STR	Doc.	-	2.6k	FREE
Milne et al. (2016)	Reach Out	SH	Doc.	1.2k	-	DUA
Mowery et al. (2016)	Twitter	DEP	Doc.	-	9.3k	AUTH
Bagroy et al. (2017)	Reddit	MHGEN	Doc.	30k	43.5k	API
De Choudhury and Kiciman (2017)	Reddit	SI	Ind.	51k	103k	API
Losada et al. (2017)	Reddit	DEP	Ind.	887	530k	DUA
Saha and De Choudhury (2017)	Reddit	STR	Doc.	-	2k	API
Shen et al. (2017)	Twitter	DEP	Ind.	300M	10B	FREE
Shen and Rudzicz (2017)	Reddit	ANX	Doc.	-	22.8k	API
Yates et al. (2017)	Reddit	DEP	Ind.	116k	-	DUA
Chancellor et al. (2018)	Reddit	EAT	Doc.	-	2.4M	API
Cohan et al. (2018)	Reddit	ANX, EAT, OCD, SCHZ, BI, PTSD, DEP, ADHD, AUT	Ind.	350k	-	DUA
Dutta et al. (2018)	Twitter	ANX	Ind.	200	209k	API
Ireland and Iserman (2018)	Reddit	ANX	Ind.	-	-	API
Li et al. (2018)	Reddit	MHGEN	Ind.	1.8k	-	API
Losada et al. (2018)	Reddit	EAT, DEP	Ind.	1.5k	1.2M	DUA
Pirina and Çöltekin (2018)	Reddit	DEP	Doc.	-	1.2k	API
Shing et al. (2018)	Reddit	SI	Ind.	1.9k	-	DUA
Sekulic et al. (2018)	Reddit	BI	Ind.	7.4k	-	API
Wolohan et al. (2018)	Reddit	DEP	Ind.	12.1k	-	API
Turcan and McKeown (2019)	Reddit	STR	Doc.	-	2.9k	FREE
Zirikly et al. (2019)	Reddit	SI	Ind.	496	32k	DUA

Table 1: Characteristics of datasets that meet our filtering criteria and are known to be accessible. Mental health conditions/predictive tasks are abbreviated as follows: Attention Deficit Hyperactivity Disorder (ADHD), Alcoholism (ALC), Anxiety (ANX), Social Anxiety (ANXS), Asperger's (ASP), Autism (AUT), Bipolar Disorder (BI), Borderline Personality Disorder (BPD), Depression (DEP), Eating Disorder (EAT), Recovery from Eating Disorder (EATR), General Mental Health Disorder (MHGEN), Obsessive Compulsive Disorder (OCD), Opiate Addiction (OPAD), Opiate Usage (OPUS), Post Traumatic Stress Disorder (PTSD), Panic Disorder (PAN), Psychosis (PSY), Trauma from Rape (RS), Schizophrenia (SCHZ), Seasonal Affective Disorder (SAD), Self Harm (SH), Stress (STR), Stressor Subjects (STRS), Suicide Attempt (SA), Suicidal Ideation (SI), Trauma (TRA).

While the majority of tasks were examined less than twice, a few tasks were considered quite frequently. Depression (42 datasets), suicidal ideation (26 datasets), and eating disorders (11 datasets) were the most common psychiatric conditions examined. Anxiety, PTSD, self-harm, bipolar disorder, and schizophrenia were also prominently featured conditions, each found within at least four unique datasets. A handful of studies sought to characterize finer-grained attributes associated with higher-level psychiatric conditions (e.g. symptoms of depression, stress events and stressor subjects) (Mowery et al., 2015; Lin et al., 2016).

Annotation. We identified 24 unique annotation mechanisms. It was common for several annotation mechanisms to be used jointly to increase precision of the defined task classes and/or evaluate the reliability of distantly supervised labeling processes. For example, some form of regular expression matching was used to construct 43 of datasets, with 23 of these including manual annotations as well. Community participation/affiliation (24 datasets), clinical surveys (22 datasets), and platform activity (3 datasets) were also common annotation mechanisms. The majority of datasets contained annotations made on the individual level (63 datasets), with the rest containing annotations made on the document level (40 datasets).[3]

Size. Of the 63 datasets with individual-level annotations, 23 associated articles described the amount of documents and 62 noted the amount of individuals available. Of the 40 datasets with document-level annotations, 37 associated articles noted the amount of documents and 12 noted the number of unique individuals. The distribution of dataset sizes was primarily left-skewed with a few notable outliers (e.g. over 100 thousand individuals) making it difficult to summarize themes using descriptive statistics.

One concerning trend that emerged across the datasets was the presence of a relatively low number of unique individuals. Indeed, these small sample sizes may further inhibit model generalization from platforms that are already demographically-skewed (Smith and Anderson, 2018). The largest datasets, which present the strongest opportunity to mitigate the issues presented by poorly representative online populations, tend to leverage the noisiest annotation mechanisms. For example, datasets that define a mainstream online community as a control group may expect to find approximately 1 in 20 of the labeled individuals are actually living with mental health conditions such as depression (Wolohan et al., 2018), while regular expressions may fail to distinguish between true and non-genuine disclosures of a mental health disorder up to 10% of the time (Cohan et al., 2018).

Primary Language. Six primary languages were found amongst the 102 datasets — English (85 datasets), Chinese (10 datasets), Japanese (4 datasets), Korean (2 datasets), Spanish (1 dataset), and Portuguese (1 dataset). This is not to say that some of the datasets do not include other languages, but rather that the predominant language found in the datasets occurs with this distribution. While an overwhelming focus on English data is a theme throughout the NLP community, it is a specific concern in this domain where culture often influences the presentation of mental health disorders (De Choudhury et al., 2017; Loveys et al., 2018).

Availability. We were able to readily identify the availability of 48 of the 102 unique datasets in our literature search using their associated articles. Of these 48 datasets, 13 were known not to be available for distribution, either due to limitations defined in the original collection agreement, removal from the public record, or ongoing analysis (Park et al., 2012; Schwartz et al., 2014; Nobles et al., 2018).

The remaining 35 datasets were available via the following distribution mechanisms (see Table 1): 18 may be reproduced with reasonable effort using an API and instructions provided within the associated article (`API`), 12 require a signed data usage agreement and/or IRB approval (`DUA`), 3 are available without restriction (`FREE`), and 2 may be retrieved directly from the author(s) with permission (`AUTH`). Of the 22 datasets that used clinically-derived annotations (e.g. mental health battery, medical history), 7 were unavailable for distribution due to terms of the original data collection process and 1 was removed from the public record. The remaining 14 had unknown availability.

4 Discussion

In this study, we introduced and analyzed a standardized directory of social media datasets used by computational scientists to model mental health phenomena. In doing so, we have provided a valuable resource poised to help researchers quickly

[3] One dataset was annotated at both a document and individual level

identify new datasets that support novel research. Moreover, we have provided evidence that affirms conclusions from Chancellor and De Choudhury (2020) and may further encourage researchers to rectify existing gaps in the data landscape. Based on this evidence, we will now discuss potential areas of improvement within the field.

Unifying Task Definitions. In just 102 datasets, we identified 24 unique annotation mechanisms used to label over 35 types of mental health phenomena. This total represents a conservative estimate given that nominally equivalent annotation procedures often varied non-trivially between datasets (e.g. PHQ-9 vs. CES-D assessments, affiliations based on Twitter followers vs. engagement with a subreddit) (Faravelli et al., 1986; Pirina and Çöltekin, 2018). Minor discrepancies in task definition reflect the heterogeneity of how several mental health conditions manifest, but also introduce difficulty contextualizing results between different studies. Moreover, many of these definitions may still fall short of capturing the nuances of mental health disorders (Arseniev-Koehler et al., 2018). As researchers look to transition computational models into the clinical setting, it is imperative they have access to standardized benchmarks that inform interpretation of predictive results in a consistent manner (Norgeot et al., 2020).

Sharing Sensitive Data. Most existing mental health datasets rely on some form of self-reporting or distinctive behavior to assign individuals into task groups, but admittedly fail to meet ideal ground truth standards. The dearth of large, shareable datasets based on actual clinical diagnoses and medical ground truth is problematic given recent research that calls into question the validity of proxy-based mental health annotations (Ernala et al., 2019; Harrigian et al., 2020). By leveraging privacy-preserving technology (e.g. blockchain, differential privacy) to share patient-generated data, researchers may ultimately be able to train more robust computational models (Elmisery and Fu, 2010; Zhu et al., 2016; Dwivedi et al., 2019). In lieu of implementing complicated technical approaches to preserve the privacy of human subjects within mental health data, researchers may instead consider establishing secure computational environments that enable collaboration amongst authenticated users (Boebert et al., 1994; Rush et al., 2019).

Addressing Bias. There remains more to be done to ensure models trained using these datasets perform consistently irrespective of population. Several studies in our review attempted to leverage demographically-matched or activity-based control groups as a comparison to individuals living with a mental health condition (Coppersmith et al., 2015b; Cohan et al., 2018). However, no study to our knowledge attempted to sample a demographically-representative cohort that would match the incidence of mental health disorders on a population level. A recent article found discrepancies between the prevalence of depression and PTSD as measured by the Centers for Disease Control and Prevention and as estimated using a model trained to detect the two conditions (Amir et al., 2019b). While the study posits reasons for the difference, it is unable to confirm any causal relationship.

The presence of downstream bias in mental health models is admittedly difficult to define and even more difficult to fully eliminate (Gonen and Goldberg, 2019; Blodgett et al., 2020). That said, the lack of demographically-representative sampling described above would serve as a valuable starting point to address. Increasingly accurate geolocation and demographic inference tools may aid in constructing datasets with demographically-representative cohorts (Huang and Carley, 2019; Wood-Doughty et al., 2020). Researchers may also consider expanding the diversity of languages in their datasets to account for variation in mental health presentation that arises due to cultural differences (De Choudhury et al., 2017; Loveys et al., 2018).

References

Silvio Amir, Mark Dredze, and John W. Ayers. 2019a. Mental health surveillance over social media with digital cohorts. In *CLPsych*.

Silvio Amir, Mark Dredze, and John W Ayers. 2019b. Mental health surveillance over social media with digital cohorts. In *CLPsych*.

American Psychiatric Association APA. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for?-a closer look at detecting mental health from language. In *CLPsych*.

John W Ayers, Benjamin M Althouse, Jon-Patrick Allem, J Niels Rosenquist, and Daniel E Ford. 2013. Seasonality in seeking mental health information on google. *American journal of preventive medicine*, 44(5):520–525.

Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. 2017. A social media based index of mental well-being in college campuses. *CHI*.

David E Bakken, R Rarameswaran, Douglas M Blough, Andy A Franz, and Ty J Palmer. 2004. Data obfuscation: Anonymity and desensitization of usable data sets. *IEEE Security & Privacy*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *First ACL Workshop on Ethics in Natural Language Processing*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp.

William E Boebert, Thomas R Markham, and Robert A Olmsted. 1994. Data enclave and trusted path system. US Patent 5,276,735.

Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*.

Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: contrasting social support around behavior change in online weight loss communities. In *CHI*.

Stevie Chancellor, Zhiyuan Lin, Erica L. Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016a. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*.

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016b. Recovery amid pro-anorexia: Analysis of recovery in social media. In *CHI*.

Dongho Choi, Ziad Matni, and Chirag Shah. 2016. What social media data should i use in my research?: A comparative analysis of twitter, youtube, reddit, and the new york times comments. In *ASIS&T*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in twitter. In *CLPsych*.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *CLPsych*.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. CLPsych 2015 shared task: Depression and PTSD on twitter. In *CLPsych*.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *CLPsych*.

Elena Davcheva, Martin Adam, and Alexander Benlian. 2019. User dynamics in mental health forums – a sentiment analysis perspective. In *Wirtschaftsinformatik*.

Munmun De Choudhury. 2015. Anorexia on tumblr: A characterization study. In *5th international conference on digital health 2015*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*.

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*.

Munmun De Choudhury, Emre Kıcıman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *CHI*.

Munmun De Choudhury, Sanket S. Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *CSCW*.

Orianna Demasi, Marti A. Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *CLPsych*.

Sarmistha Dutta, Jennifer Ma, and Munmun De Choudhury. 2018. Measuring the impact of anxiety on online social interactions. In *ICWSM*, pages 584–587.

Ashutosh Dhar Dwivedi, Gautam Srivastava, Shalini Dhar, and Rajani Singh. 2019. A decentralized privacy-preserving healthcare blockchain for iot. *Sensors*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*.

Ahmed M Elmisery and Huaiguo Fu. 2010. Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. In *2010 IEEE 34th Annual Computer Software and Applications Conference Workshops*.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *CHI*.

Carlo Faravelli, Giorgio Albanesi, and Enrico Poli. 1986. Assessment of depression: a comparison of rating scales. *Journal of affective disorders*.

Christian Fuchs. 2015. *Culture and economy in the age of social media*. Routledge.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *CLPsych*.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.

Marc H Gorelick. 2006. Bias arising from missing data in predictive models. *Journal of clinical epidemiology*.

Melissa W Graham, Elizabeth J Avery, and Sejin Park. 2015. The role of social media in local government crisis communications. *Public Relations Review*.

Jonathan Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David R. Traum, Albert A. Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *"Findings of ACL: EMNLP"*.

Eben Holderness, Philip Cawkwell, Kirsten Bolton, James Pustejovsky, and Mei-Hua Hall. 2019. Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. In *ClinicalNLP*.

Binxuan Huang and Kathleen M Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation.

Molly Ireland and Micah Iserman. 2018. Within and between-person differences in language used across anxiety support and neutral reddit communities. In *CLPsych*.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *CLPsych*.

Jared Jashinsky, Scott H. Burton, Carl Lee Hanson, Joshua H. West, Christophe G. Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*, 35 1:51–9.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. *HT*.

Yaoyiran Li, Rada Mihalcea, and Steven R. Wilson. 2018. Text-based detection and understanding of changes in mental health. In *SocInfo*.

Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *22nd ACM international conference on Multimedia*.

Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua. 2016. What does social media say about your stress?. In *IJCAI*, pages 3775–3781.

David E Losada, Fabio Crestani, and Javier Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *CLEF*.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: early risk prediction on the internet. In *CLEF*.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *CLPsych*.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. In *CLPsych*.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *CLPsych*.

Danielle Mowery, Craig Bryan, and Mike Conway. 2015. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In *CLPsych*.

Danielle L. Mowery, Albert Park, Craig J Bryan, and Mike Conway. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. In *PEOPLES*.

Alicia L. Nobles, Jeffrey J. Glenn, Kamran Kowsari, Bethany A. Teachman, and Laura E. Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. *CHI*.

Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. 2020. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nature medicine*.

Galen Panger. 2016. Reassessing the facebook experiment: critical thinking about the validity of big data research. *Information, Communication & Society*, 19(8):1108–1126.

Minsu Park, Chiyoung Cha, and Meeyoung Cha. 2012. Depressive moods of users portrayed in twitter.

A Perrin and M Anderson. 2019. Share of us adults using social media, including facebook, is mostly unchanged since 2018. pew research center.

Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *SMM4H*.

W Nicholson Price and I Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature medicine*.

Brenna N Renn, Abhishek Pratap, David C Atkins, Sean D Mooney, and Patricia A Areán. 2018. Smartphone-based passive assessment of mobility in depression: Challenges and opportunities. *Mental health and physical activity*, 14:136–139.

Sarah Rush, Sara Britt, and John Marcotte. 2019. Icpsr virtual data enclave as a collaboratory for team science.

Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *CSCW*.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through Facebook. In *CLPsych*.

Ivan Sekulic, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. In *9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *CLPsych*.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *CLPsych*.

Aaron Smith and Monica Anderson. 2018. Social media use in 2018. *Pew*.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *LOUHI*.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *LCCM Workshop*.

Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *JMIR*, 19(6):e228.

Zach Wood-Doughty, Paiheng Xu, Xiao Liu, and Mark Dredze. 2020. Using noisy self-reports to predict twitter user demographics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *EMNLP*.

Lei Zhang, Xiaolei Huang, Tianli Liu, Zhenxiang Chen, and Tingshao Zhu. 2014. Using linguistic features to estimate suicide probability of chinese microblog users. In *HCC*.

Haining Zhu, Joanna Colgan, Madhu Reddy, and Eun Kyoung Choe. 2016. Sharing patient-generated data in clinical practices: an interview study. In *AMIA*.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *CLPsych*.