On the State of Social Media Data for Mental Health Research

Abstract

Data-driven research for mental health that leverages social media has become a major focus in computational science research. However, progress in the domain, in terms of medical understanding and system performance, remains bounded by the availability of adequate data. Prior systematic reviews have not necessarily made it possible to infer the degree to which data-related challenges have affected research progress. In this paper, we offer an analysis 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 data sets, annotated using a standardized schema to facilitate meta-analysis.

Introduction

Numerous studies analyzing online data, including social media, have yielded new insights into population-level mental health and shown promising avenues for the incorporation of data-driven analyses in the treatment of psychiatric disorders. Success of any data-driven research agenda requires access to large amounts of relevant data. Social media in support of mental health research typically includes both raw and annotated data collected from a combination of publicly-accessible interfaces and traditional human subjects studies.

Obtaining a sufficient sample size of high-quality data to examine mental health is often more complex than it is in other areas of research that use social media. Considerations specific to the mental health space necessitate extra care when creating and supporting new data sets. For instance, behavioral disorders are known to display variable clinical presentations amongst different populations, which makes identification of ground truth for learning purposes quite difficult (Choudhury et al. 2017; Arseniev-Koehler, Mozgai, and Scherer 2018). Common methods for capturing an individual's diagnosis, such as using regular expressions to identify self-reported diagnoses or grouping individuals based on activity patterns, provide opportunities for constructing large-scale data sets (Coppersmith et al. 2015a; Kumar et al. 2015). Still, they rely on stereotypical hypotheses that lack the same clinical confidence as something like a mental health battery (Zhang et al. 2014).

Furthermore, even with methods for identifying cohorts of individuals living with a mental health disorder, ethical considerations complicate large-scale sharing of such sensitive data (Benton, Coppersmith, and Dredze 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 difficult to implement and are harder to apply when clinical information is involved due to HIPAA. Moreover, many privacy-preserving practices require that potentially useful signal, such as an individual's demographics or their social network, are discarded. In addition to hindering predictive performance, this missingness has the potential to limit fairness, generalizability, and reproducibility.

A healthy research community requires access to high quality data. While prior systematic reviews of computational research for mental health, such as those from Guntuku et al. (2017) and Wongkoblap, Vadillo, and Curcin (2017), have made research in this domain more accessible, the state of data in support of this research remains unclear. This paper offers an analysis on the state of social media data that exists for conducting mental health research. We support this analysis with an open-source directory of mental-health data sets,[1] annotated using a standardized schema to facilitate meta-analysis. We use this directory to quantify shortcomings of current mental health data sets and thereafter offer recommendations for future work.

Criteria for Data Set Selection

To create a directory of data sets, we establish selection criteria and a list of data set attributes we will collect and analyze. We build upon two previously conducted systematic reviews of the mental health landscape, adding recent work and expanding the types of mental-health disorders discussed (Guntuku et al. 2017; Wongkoblap, Vadillo, and Curcin 2017). In our selection of data sets we emphasize both relevance to psychiatric diagnosis and data availability.

[1] www.repository-available-after-anonymity-period.com

Data Set Identification. Data sets were sourced using a breadth-focused literature search, though we cannot claim that the resulting enumeration is fully comprehensive. After including data sources from the aforementioned systematic reviews (Guntuku et al. 2017; Wongkoblap, Vadillo, and Curcin 2017), we searched for relevant literature that lie primarily at the intersection of natural language processing (NLP) and mental health communities. We sought peer-reviewed studies published after 2012 in related conferences (e.g. *NAACL*, *EMNLP*, *ACL*, *COLING*), workshops (e.g. *CLPsych*, *LOUHI*), and health-focused journals (e.g. *PNAS*, *BMJ*).

We searched *Google Scholar*, *ArXiv*, and *PubMed* to identify additional candidate articles. We primarily used two structures of search terms: 1) *(mental health | MH-DISORDER) + (social | electronic) + media* and 2) *(machine learning | prediction | inference | detection) + (mental health | MH-DISORDER)*, where and "|" indicates a logical OR and *MH-DISORDER* was one of 13 mental health keywords.[2] Additional literature was identified using snowball sampling from the citations of these papers. To restrict the scope of this work, computational research regarding neurodegenerative disorders (e.g. Dementia, Parkinson's Disease) were excluded at this time.

Selection Criteria Our first criteria was that data sets must be based on non-clinical electronic media sources (e.g. social media, SMS). For example, we exclude data sets derived solely from electronic health records or personal interviews where there is no accompanying electronic media (Gratch et al. 2014; Holderness et al. 2019). While excluded from this paper, we maintain annotated references to data sets that lie tangential to this criteria in the aforementioned digital directory.

Our second criteria requires data sets to contain a dependent variable that captures or proxies a well-defined mental health disorder. This excludes data on date of diagnosis, cyberbulling, and depression framed as a sentiment analysis task (Sui 2015; MacAvaney et al. 2018; Davcheva, Adam, and Benlian 2019). We also exclude data sets 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, Hearst, and Recht 2019)).

Annotation Schema. We develop a high-level schema to code properties of each data set discovered in the literature review. In addition to standard reference information (e.g. Title, Year, 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, Clinical Diagnoses)
- **Annotation Level**: Resolution at which the 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 data set
- **Data Availability**: Whether the data set can be shared and, if so, the mechanism by which it may be accessed (e.g. Data Usage Agreement (DUA), Reproducible via API, Distribution Prohibited by Collection Agreement)

If a characteristic is not clear from the data set'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 more details in the directory.

Analysis

Our literature search yielded 71 articles referencing 58 nominally-unique data sets published between 2012 and 2019. The inclusion/exclusion criteria left us with 44 data sets. A majority of the data sets were released after 2014, with an average of 5.5 per year, a minimum of 1 (2012) and a maximum of 8 (2018). The 2015 CLPsych Shared Task (Coppersmith et al. 2015a) and Reddit Self-reported Depression Diagnosis (Yates, Cohan, and Goharian 2017) data sets were the most reused resources, serving as the basis of 6 and 2 publications respectively. Other reused data sets were identified in no more than one additional paper each. Table 1 lists all known available data sets along with some of our annotations. Remaining data sets with similar annotations may be found in our digital directory, alongside code that performs filtering based on the selection criteria.

Platforms. We identified 12 unique electronic media platforms across the 44 data sets. Twitter (19 data sets) and Reddit (12) were the most widely used platforms. We did not encounter a data set that captures an individual's behavior across multiple platforms.

Despite being the three most-widely adopted social media platforms (Perrin and Anderson 2019), YouTube, Facebook, and Instagram were relatively underutilized for mental health research, each found less than three times in our analysis. We expect our focus on NLP to moderate the presence of YouTube and Instagram based data sets, 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. 2016; Choi, Matni, and Shah 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, Coppersmith, and Dredze 2017).

Tasks. We identified 28 unique mental-health related modeling tasks across the 44 data sets. While the majority of tasks were examined less than twice, a few tasks were frequent. Depression (19 data sets) and suicidal ideation (17) were common. Modeling of PTSD, Bipolar Disorder, Self-harm, and Eating Disorders were also prominent tasks, each found within at least four unique data sets.

Annotation. We identified 14 unique types of annotation mechanisms. It was common for several annotation mech-

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

Reference	Platform(s)	Task(s)	Level	# Inds.	# Docs.	Availability
Jashinsky et al. (2014)	Twitter	SI	Doc.	594k	733k	API
Coppersmith, Dredze, and Harman (2014)	Twitter	BIPD, PTSD, SAD, DEP	Ind.	7k	16.7M	DUA
Coppersmith, Harman, and Dredze (2014)	Twitter	PTSD	Ind.	6.3k		DUA
Kumar et al. (2015)	Wikipedia, Reddit	SI	Ind.	66k	19.1k	API
Mowery, Bryan, and Conway (2015)	Twitter	DEP	Doc.		129	Contact Author
Coppersmith et al. (2015a)	Twitter	PTSD, DEP	Ind.	1.7k		DUA
Coppersmith et al. (2015b)	Twitter	ANX, EAT, OCD, SCHZ, SAD, BIPD, PTSD, DEP, ADHD	Ind.	1.9k	6.4M	DUA
Choudhury et al. (2016)	Reddit	PSY, EAT, ANXS, SH, BIPD, PTSD, RS, DEP, PAN, SI, TRA	Ind.	880		API
Mowery et al. (2016)	Twitter	DEP	Doc.		9.3k	Contact Author
Coppersmith et al. (2016)	Twitter	SA	Ind.	250		DUA
Gkotsis et al. (2016)	Reddit	ANX, BRPD, SCHZ, SH, ALC, BIPD, OPAD, ASP, SI, AUT, OPUS	Ind.			API
Milne et al. (2016)	Reach Out	SH	Doc.	1.2k		DUA
Bagroy, Kumaraguru, and Choudhury (2017)	Reddit	MHGEN	Doc.	30k	43.5k	API
Yates, Cohan, and Goharian (2017)	Reddit	DEP	Ind.	116k		DUA
Choudhury and Kiciman (2017)	Reddit	SI	Ind.	103k	53.3k	API
Wolohan et al. (2018)	Reddit	DEP	Ind.	12.1k		API
Li, Mihalcea, and Wilson (2018)	Reddit	MHGEN	Ind.	1.8k		API
Shing et al. (2018)	Reddit	SI	Ind.	1.9k		DUA
Cohan et al. (2018)	Reddit	ANX, EAT, OCD, SCHZ, BIPD, PTSD, DEP, ADHD, AUT	Ind.	49.2k		DUA
Turcan and McKeown (2019)	Reddit	STR	Doc.		2.9k	Freely
Zirikly et al. (2019)	Reddit	SI	Ind.	496	32k	DUA

Table 1: Characteristics of data sets that both meet our filtering criteria and are known to be accessible. Mental health disorders/predictive tasks are abbreviated as follows: ADHD (ADHD), Alcoholism (ALC), Anxiety (ANX), Social Anxiety (ANXS), Aspergers (ASP), Autism (AUT), Bipolar Disorder (BIPD), Borderline Personality Disorder (BRPD), Depression (DEP), Eating Disorer (EAT), General Mental Health Disorder (MHGEN), OCD (OCD), Opiate Addiction (OPAD), Opiate Usage (OPUS), PTSD (PTSD), Panic Disorder (PAN), Psychosis (PSY), Trauma from Rape (RS), Schizophrenia (SCHZ), Seasonal Affective Disorder (SAD), Self Harm (SH), Stress (STR), Suicide Attempt (SA), Suicidal Ideation (SI), Trauma (TRA).

anisms to be used jointly to either increase precision of the defined task classes or evaluate the reliability of distantly supervised labeling processes. For example, some form of regular expression matching was used in 18 of studies, with 9 of these also using manual annotation to either discard false positives or quantify the label precision on a subsample of the larger data set. Clinical surveys (9 studies), community-participation (7), and platform activity (3) were also common annotation mechanisms. The majority of data sets contained annotations per individual (29), with the rest containing an annotation per document (15).

Size. Of the 15 data sets with document-level annotations, 12 associated articles noted the amount of documents and 7 noted the number of unique individuals. Likewise, of the 29 data sets with individual-level annotations, 8 articles described the amount of documents and 28 noted the amount of individuals available. The distribution of data set sizes was primarily left-skewed with a few notable outliers that make descriptive statistics unreliable for reporting.

One concerning theme across the data sets was a relatively low number of unique individuals; the small sample size may further inhibit generalization from already demographically non-representative platforms (Smith and Anderson 2018). The largest data sets, which in some cases have millions of documents and individuals, tend to leverage regular expressions or community participation to annotate data points. While these approaches mitigate the issue of having a non-representative sample, they may also introduce noise. For example, data sets that define a mainstream online community as a control group may find approximately 1 in 20 of the labeled individuals are actually living with depression (Wolohan et al. 2018). Similarly, even high-precision regular expressions may fail to distinguish between true and non-genuine disclosures of a mental-health disorder anywhere from 4.2% to 10% of the time (Cohan et al. 2018).

Primary Language. Four unique primary languages were

found within the 44 data sets—English (38), Chinese (4), Korean (1), and Japanese (1). This is not to say that some of the data sets do not include other languages, but rather that the predominant language found in the data sets occurs with this distribution. While the overwhelming focus on English is a common theme through much of the NLP community, it is of potential concern in this domain where cultural expectations and standards often motivate different presentations of mental health disorders (Choudhury et al. 2017; Loveys et al. 2018).

Availability. We were able to identify the availability of 27 of the 44 unique data sets in our literature search. Of these 27, 5 were known not to be available for distribution, either due to limitations defined in the original collection agreement or removal from the public record. One data set may become available after proprietary analysis.

The remaining 21 data sets were available via the following distribution mechanisms: 10 require a signed data usage agreement (DUA) and/or IRB approval, 8 may be reproduced with reasonable effort using an API and instructions within the associated article, 2 can be retrieved from the author(s) with permission, and 1 may be retrieved without restriction using a public download link.

Of the 9 data sets that used clinically-derived annotations (e.g. mental health battery, medical history), 2 were unavailable for distribution due to terms of the data collection process. The remaining 7 had unknown availability. 4 of these data sets were non-English, which we did not attempt to access during this pilot study. An additional 2 data sets were not related to depression or suicidal ideation and were also not requested as part of this study.

Discussion

We introduced and analyzed a standardized directory of data sets used to model several mental health conditions. This directory will aid in researchers identifying existing data sets to support research, as well as clarify gaps in the data landscape. Unlike prior literature reviews, our attention toward data instead of computational modeling approaches has illuminated existing shortcomings in the research domain. In particular, we noted two primary weaknesses amongst available data sets for mental health research: 1) ambiguous task definitions and 2) susceptibility to bias.

With respect to the first weakness: in 44 data sets we found 28 unique modeling tasks serving as a proxy for the detection of mental health conditions. Although annotation techniques are often similar, there remains ample variance in the way even a single mental health condition is defined. While minor discrepancies in task definition reflect the heterogeneity of how several mental health conditions are manifested, they may also introduce difficulty contextualizing results between different studies and still fall short of capturing the nuances of mental health disorders (Arseniev-Koehler, Mozgai, and Scherer 2018).

The dearth of data sets based on clinical-diagnoses and medical ground truth further highlights the ambiguity in existing task definitions. Most existing mental health data sets 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. Finding privacy-preserving means to share patient-generated data may have a valuable impact on resolving the ambiguity currently found in mental health data sets (Zhu et al. 2016).

Despite the variety of task definitions, there remains more to be done to ensure models trained using these data sets perform consistently for all demographic groups. Several studies attempted to leverage demographically-similar or activity-based control groups as a comparison to individuals living with a mental health condition (Coppersmith et al. 2015a; 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, Dredze, and Ayers 2019). While the study posits reasons for the difference, it is unable to confirm any causal relationship.

As seen in a variety of other NLP tasks, the presence of downstream bias in mental health models is admittedly difficult to fully eliminate. That said, the lack of demographically-representative sampling described above would serve as a valuable starting point to address. Additionally, researchers may consider expanding the scope of languages their data sets encompass. Increasingly accurate demographic and geolocation inference tools may aid in constructing data sets with large-scale, demographically-representative cohorts. The creation of such data sets may further clarify cultural nuances of mental health manifestation (Loveys et al. 2018).

References

Amir, S.; Dredze, M.; and Ayers, J. W. 2019. Mental health surveillance over social media with digital cohorts. In *CLPsych*.

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

Bagroy, S.; Kumaraguru, P.; and Choudhury, M. D. 2017. A social media based index of mental well-being in college campuses. *CHI*.

Benton, A.; Coppersmith, G.; and Dredze, M. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102.

Chancellor, S.; Lin, Z.; Goodman, E. L.; Zerwas, S.; and Choudhury, M. D. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*.

Choi, D.; Matni, Z.; and Shah, C. 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*.

Choudhury, M. D., and Kiciman, E. 2017. The language

of social support in social media and its effect on suicidal ideation risk. In *ICWSM*.

Choudhury, M. D.; Kıcıman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *CHI*.

Choudhury, M. D.; Sharma, S. S.; Logar, T.; Eekhout, W.; and Nielsen, R. C. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *CSCW*.

Cohan, A.; Desmet, B.; Yates, A.; Soldaini, L.; MacAvaney, S.; and Goharian, N. 2018. Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions. In *COLING*.

Coppersmith, G.; Dredze, M.; Harman, C.; Hollingshead, K.; and Mitchell, M. 2015a. CLPsych 2015 shared task: Depression and PTSD on twitter. In *CLPsych*.

Coppersmith, G.; Dredze, M.; Harman, C.; and Hollingshead, K. 2015b. From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *CLPsych*.

Coppersmith, G.; Ngo, K.; Leary, R.; and Wood, A. 2016. Exploratory analysis of social media prior to a suicide attempt. In *CLPsych*.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in twitter. In *CLPsych*.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Davcheva, E.; Adam, M.; and Benlian, A. 2019. User dynamics in mental health forums – a sentiment analysis perspective. In *Wirtschaftsinformatik*.

Demasi, O.; Hearst, M. A.; and Recht, B. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *CLPsych*.

Gkotsis, G.; Oellrich, A.; Hubbard, T.; Dobson, R.; Liakata, M.; Velupillai, S.; and Dutta, R. 2016. The language of mental health problems in social media. In *CLPsych*.

Gratch, J.; Artstein, R.; Lucas, G. M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; Traum, D. R.; Rizzo, A. A.; and Morency, L.-P. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*.

Guntuku, S. C.; Yaden, D. B.; Kern, M. L.; Ungar, L. H.; and Eichstaedt, J. C. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18:43–49.

Holderness, E.; Cawkwell, P.; Bolton, K.; Pustejovsky, J.; and Hall, M.-H. 2019. Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. In *ClinicalNLP*.

Jashinsky, J.; Burton, S. H.; Hanson, C. L.; West, J. H.; Giraud-Carrier, C. G.; Barnes, M. D.; and Argyle, T. 2014. Tracking suicide risk factors through twitter in the us. *Crisis* 35 1:51–9.

Kumar, M.; Dredze, M.; Coppersmith, G.; and Choudhury, M. D. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. *HT*.

Li, Y.; Mihalcea, R.; and Wilson, S. R. 2018. Text-based detection and understanding of changes in mental health. In *SocInfo*.

Loveys, K.; Torrez, J.; Fine, A.; Moriarty, G.; and Coppersmith, G. 2018. Cross-cultural differences in language markers of depression online. In *CLPsych*.

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 *CLPsych*.

Milne, D. N.; Pink, G.; Hachey, B.; and Calvo, R. A. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *CLPsych*.

Mowery, D. L.; Park, A.; Bryan, C. J.; and Conway, M. 2016. Towards automatically classifying depressive symptoms from twitter data for population health. In *PEOPLES*.

Mowery, D.; Bryan, C.; and Conway, M. 2015. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In *CLPsych*.

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

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

Shing, H.-C.; Nair, S.; Zirikly, A.; Friedenberg, M.; Daumé, H.; and Resnik, P. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *CLPsych*.

Smith, A., and Anderson, M. 2018. Social media use in 2018. *Pew*.

Sui, J. 2015. Understanding and fighting bullying with machine learning.

Turcan, E., and McKeown, K. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *LOUHI*.

Wolohan, J.; Hiraga, M.; Mukherjee, A.; Sayyed, Z. A.; and Millard, M. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *LCCM Workshop*.

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

Yates, A.; Cohan, A.; and Goharian, N. 2017. Depression and self-harm risk assessment in online forums. In *EMNLP*.

Zhang, L.; Huang, X.; Liu, T.; Chen, Z.; and Zhu, T. 2014. Using linguistic features to estimate suicide probability of chinese microblog users. In *HCC*.

Zhu, H.; Colgan, J.; Reddy, M.; and Choe, E. K. 2016. Sharing patient-generated data in clinical practices: an interview study. In *AMIA*.

Zirikly, A.; Resnik, P.; Uzuner, Ö.; and Hollingshead, K. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *CLPsych*.