Improving Mental Health Classifier Generalization with Pre-diagnosis Data

Yujian Liu[*†1], **Laura Biester**[*2], **Rada Mihalcea**[2]
[1] University of California, Santa Barbara
[2] University of Michigan
yujianliu@ucsb.edu, {lbiester, mihalcea}@umich.edu

Abstract

Recent work has shown that classifiers for depression detection often fail to generalize to new datasets. Most NLP models for this task are built on datasets that use textual reports of a depression diagnosis (e.g., statements on social media) to identify diagnosed users; this approach allows for collection of large-scale datasets, but leads to poor generalization to out-of-domain data. Notably, models tend to capture features that typify direct discussion of mental health rather than more subtle indications of depression symptoms. In this paper, we explore the hypothesis that building classifiers using exclusively social media posts from before a user's diagnosis will lead to less reliance on shortcuts and better generalization. We test our classifiers on a dataset that is based on an external survey rather than textual self-reports, and find that using pre-diagnosis data for training yields improved performance with many types of classifiers.

1 Introduction

In recent years, computational methods, including Natural Language Processing (NLP), have been applied to social media data with the objective of learning about mental illness and improving mental healthcare (e.g., Coppersmith et al. 2015; Mitchell, Hollingshead, and Coppersmith 2015; Jamil et al. 2017; Cohen et al. 2020). A significant amount of work in this area focuses on the task of predicting mental health status from social media content. The main signals that have been used in order to infer mental health status for these classification tasks are listed in Table 1.

The practices of using self-reported diagnoses and community membership show promise from a machine learning perspective, in that data can be automatically labeled, and collecting datasets does not require participation from study "participants" in the form of surveys. This allows large datasets to be collected, which lend themselves well to deep learning methods. However, recent work has questioned the validity of these methods and their ability to generalize to new populations (Harrigian, Aguirre, and Dredze 2020; Ernala et al. 2019). Self-report bias has been identified as a major drawback of much of the work on mental health in machine learning (Chancellor et al. 2019). This leads to the key question we address in this paper: is it possible to take advantage of the benefits afforded by automatic labeling practices (e.g., easy, unobtrusive data collection) while decreasing the generalization issues?

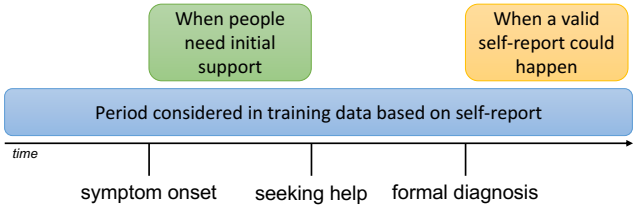

Figure 1: By including all data from people who self-report, we introduce a mismatch between our training data and some real-life use cases. This problem can be reduced by modifying the time period considered in the training data.

If our goals include early intervention and population-level monitoring, we should aim to do well at classifying all users who are symptomatic. The people included in datasets based on self-report differ from those who have undiagnosed depression in that they have sought help and received a diagnosis from a professional. Their higher likelihood to be receiving treatment makes them differ from target populations in substantial ways, which may mean that features learned by classifiers will not generalize to all of those who are symptomatic. This mismatch means that these classifiers may not identify those who would most benefit from being connected to support (Figure 1).

In this paper, we explore whether using data from before users are diagnosed with depression (pre-diagnosis data) can improve generalization to populations who do not *by definition* discuss mental health online. We take advantage of the observation that there was a period of time during which users with a self-reported diagnosis were not yet diagnosed, and may not have posted explicitly mental health-related content on social media. Prior work has shown significant changes in user behavior after reporting a schizophrenia diagnosis on Twitter (Ernala et al. 2017); these changes are attributed to the therapeutic benefits of self-disclosure, but could also be linked to treatment. Furthermore, before being diagnosed, people will have some symptoms of depres-

[*]These authors contributed equally.
[†]Work completed as a student at the University of Michigan.

Method	A person is labeled as depressed if...	Examples
Surveys and Healthcare Collaborations	their PHQ9 score on a survey reaches the level required to diagnose clinical depression.	De Choudhury et al. (2013); Ernala et al. (2019)
Self-Reported Diagnosis	they post "I have been diagnosed with depression" (or similar) on social media.	Coppersmith, Dredze, and Harman (2014); Cohan et al. (2018)
Community Membership	they join the r/depression Reddit community	Shen and Rudzicz (2017); Wolohan et al. (2018)

Table 1: Signals used to infer mental health status for classification tasks. The lists of methods used for labeling and example papers are not exhaustive.

sion, but they may be less likely to outwardly discuss mental health. By using data exclusively from the pre-diagnosis stage for training, we hypothesize that we may be able to build classifiers that do not take advantage of *shortcuts* that lead to poor generalization.

To explore this hypothesis, we collect a dataset based on self-reported depression diagnoses; then, we extract the diagnosis timestamp for a subset of users. Finally, we build classifiers and test them on a dataset from another social media platform where mental health status is determined based on an external survey, rather than people's behavior on the platform.[1] We find that for some types of classifiers, generalization to the new population improves when using pre-diagnosis data for training.

2 Related Work

Mental health related textual data has been difficult to collect due to privacy issues and the cost of diagnosis. However, with the explosion of user-generated social media content, researchers have begun to consider how such content can be leveraged for the analysis of language usage related to mental health.

Prior work has mainly adopted two proxy signals to identify people with mental health conditions on social media platforms. The first and most popular proxy signal is a set of self-reported diagnosis patterns. Coppersmith, Dredze, and Harman (2014) used patterns like "I was diagnosed with X" to identify more than 1,200 Twitter users with four mental health conditions (depression, bipolar disorder, PTSD, and SAD). Following this work, similar patterns were created for additional mental health conditions and different social media platforms (Coppersmith et al. 2015; Mitchell, Hollingshead, and Coppersmith 2015; Ernala et al. 2017; Yates, Cohan, and Goharian 2017; Cohan et al. 2018; Birnbaum et al. 2017). Based on these diagnosis patterns, some work further includes experts to verify the authenticity of identified diagnosed users (Mitchell, Hollingshead, and Coppersmith 2015; Ernala et al. 2017; Cohan et al. 2018; Birnbaum et al. 2017). The second proxy signal is the communities that users affiliate themselves with. McManus et al. (2015) identifies individuals with schizophrenia by checking if they follow the Twitter account @schizotribe. Jamil et al. (2017) identifies users who may have depression by searching within the #BellLetsTalk campaign. Similarly, affiliation behaviors on Reddit have also been used to identify individuals with mental health conditions. Participation in subreddits such as r/Anxiety and r/SuicideWatch are used as proxy signals to identify people with mental health conditions (Gkotsis et al. 2017; Shen and Rudzicz 2017).

While identifying diagnosed people through proxy signals, some work analyzed the amount of time that passes between diagnosis and self-report (MacAvaney et al. 2018). However, they focus on classifying diagnosis recency and condition state, and do not study the impact of the time period spanned by user's data on classifiers that predict mental health conditions. Eichstaedt et al. (2018) study how using data from various time periods before diagnosis affects classifier performance; however, they address a fundamentally different question than we do, focusing on *how early* depression can be identified, rather than classifier generalizability, and on periods only before diagnosis. Similarly, De Choudhury et al. (2013) and Losada, Crestani, and Parapar (2018) focus on early detection of depression. Uban, Chulvi, and Rosso (2021) look at how the evolution of emotional and cognitive processing language differs between depressed and control groups over time, along with building a classifier with a hierarchical attention network based on the knowledge that *change* in one's mental state is important for depression classification.

Although social media platforms provide convenient access to a large amount of mental health data, previous work has identified several pitfalls when using such proxy signals to identify people with mental health conditions. Ernala et al. (2019) shows that people identified by proxy signals have different behaviors than people who are clinically diagnosed but do not post about mental health on social media. As a result, machine learning classifiers trained on such proxy signals cannot generalize to other populations that do not talk about mental health on social media. Harrigian, Aguirre, and Dredze (2020) founds that models trained on data collected using a variety of proxy signals do not generalize across different social media platforms and proxies.

3 Data

In this section, we describe our method for collecting two datasets for the analysis of linguistic classification based on people's mental health diagnoses. Specifically, we focus on English language user generated content on Reddit (§3.1) and Twitter (§3.2), and we analyze users with depressive disorders (depression). Following Cohan et al. (2018) whose

[1]Code: https://github.com/MichiganNLP/prediagnosis

	# users	# posts per user	post length	# MH posts per user	MH post length
Diagnosed users	20,573	753.8 (±1221.5)	40.0 (±76.6)	51.9 (±116.2)	102.9 (±166.9)
Control users	185,157	497.3 (±957.2)	18.9 (±45.6)	—	—

Table 2: Statistics of SELFREPORT training set, which is based on self-reported depression diagnoses. Post length is measured in tokens.

	# users	# tweets per user	tweet length
diagnosed-depression	32	1696.9 (±1481.7)	12.8 (±7.6)
diagnosed-all	55	2974.8 (±9948.8)	7.3 (±7.3)
control	138	1515.3 (±3913.0)	9.2 (±7.5)

Table 3: Twitter SURVEY-based dataset statistics. Post length is measured in tokens. *diagnosed-all* is a superset of *diagnosed-depression*; there are a total of 193 users included in the SURVEY-based dataset.

		diagnosed-depression ($n = 32$)	*diagnosed-all* ($n = 55$)	*control* ($n = 138$)
gender	Female	20	33	63
	Male	10	20	73
	Self-identify	0	0	2
	Genderqueer/Gender non-conforming	1	1	0
	Trans male/Trans man	1	1	0
race	White	22	40	102
	Two or More Races	6	9	15
	Asian American / Asian	0	1	10
	African American / Black	0	0	8
	Hispanic / Latino/a	4	5	2
	Self Identify/Unreported	0	0	1
class year	Freshman	8	15	48
	Sophomore	7	12	40
	Junior	8	13	26
	Senior	8	14	23
	Other	1	1	1
school	Liberal Arts and Sciences	16	29	63
	Other	11	19	37
	Engineering	4	5	30
	Business	1	2	8

Table 4: Twitter SURVEY-based dataset user demographics. Demographics are reported for all users in our study who shared their Twitter handles. For brevity, respondents who listed multiple races in their responses are listed as "Two or More Races" and respondents outside of the three largest schools at the university are listed as "Other".

data collection process we built on, we do not release raw Reddit data, but share code for data collection. We cannot release the Twitter data due to the possibility of identifying individuals in the dataset (who may not have publicly shared their diagnosis) by searching for the text in their tweets.

3.1 Reddit Self-Report-Based Dataset (SELFREPORT)

Data collection. We follow Cohan et al. (2018) by using self-reported diagnosis patterns to identify diagnosed users and collect corresponding control users based on their activity on Reddit. We look at all submissions and comments on Reddit from January 2006 to December 2019 using PushShift (Baumgartner et al. 2020). For convenience, we will use the term "post" to refer to both submissions and comments hereafter, and we do not distinguish them. Our dataset expands the Self-reported Mental Health Diagnoses (SMHD) dataset (Cohan et al. 2018) in three ways. First, we collect data from a longer time period. Second, we expand the list of mental health related keywords and subreddits used in SMHD. Third, our dataset includes self-report posts, which allow us to extract a diagnosis time (§4) and identify pre-diagnosis posts.

Diagnosed users are identified using an existing list of self-reported diagnosis patterns from SMHD.[2] An example of such a pattern is "I have been diagnosed with depression." To reduce the false positive rate in retrieved data, another list of negative diagnosis patterns, e.g., "I'm not technically diagnosed," is used to remove users who do not have depression but are retrieved by the diagnosis patterns. The patterns are crafted using terms like "diagnosed" and "clinically" to capture mentions of *clinical depression*, rather than more colloquial uses of the word "depressed," e.g., "I am depressed because the football team is not doing well." Using such positive and negative patterns results in a set of diagnosed users with high precision (Cohan et al. 2018); all labeling was done automatically, without human annotation. After identifying users who make a post that matches these patterns (in any subreddit), we collect their posts across *all subreddits*. Finally, we remove users who do not have at least 50 posts that are not about mental health, following the pro-

[2]https://ir.cs.georgetown.edu/data/smhd/

cedure outlined at the end of this section. The choice of 50 follows the practice from used in prior work (Cohan et al. 2018); this filtering is done in order to ensure that there is enough data for each user to train our classifiers.

Control users are identified for each diagnosed user based on their post activity. Specifically, we use three conditions for finding control users: (1) each control user must post in at least one common subreddit with the diagnosed user; (2) the number of posts of each paired control user and diagnosed user cannot deviate by a factor larger than two; and (3) control users cannot have any mental health related posts. For each diagnosed user, we find nine corresponding control users.

Mental health related data. We follow Cohan et al. (2018) by using a set of mental health terms and mental health subreddits to identify posts that relate to mental health. The posts including these terms are excluded when training classifiers to allow for better generalizability. In preliminary experiments, we found that the existing list does not include some terms and subreddits that closely relate to mental health (e.g., names of antidepressants and r/2meirl4meirl[3]). Antidepressants and terms posted in r/2meirl4meirl tended to have high weights in linear classifiers, but using them for classification does not generalize to a population that does not explicitly talk about mental health. We therefore extend the existing list by adding a list of common antidepressants[4] and additional mental health related subreddits.

Data statistics. We collect 29,390 diagnosed users and 264,510 control users in total. We randomly split our dataset by user into train, validation, and test sets, which contain 70%, 15%, and 15% of the users, respectively. We present the statistics of the training set in Table 2. We observe that diagnosed users tend to have more and longer posts than control users. Furthermore, for the same set of diagnosed users, mental health related posts (which are excluded when training and testing) tend to be longer than other posts.

Data filtering and cleaning. We identify and remove users who appeared to be bots from our dataset. We found that bots tended to have a very high number of posts and either explicitly stated that they were bots or used extremely repetitive language (Massanari 2016). We manually examine posts from users with a large number of posts, and remove them if they appear to be bots. We clean the text by removing special characters and sequences, such as newlines, quotes, emails, and tables, as has been done in prior work using social media data (e.g., Campillo-Ageitos, Martinez-Romo, and Araujo (2022); Mukherjee and Das (2022)).

3.2 Twitter Survey-Based Dataset (SURVEY)

In order to test our classifiers on a dataset that is not built on proxy signals, we use a dataset from Twitter. The dataset includes the Twitter handles from 210 students at a large US university.[5] Tweets are scraped using the Tweepy library.[6] The students provided their Twitter handles in 2018 and 2019, and also completed a survey asking if the student had been diagnosed with a mental health condition. We split the students into three sets: students who state that they have diagnosed depression (*diagnosed-depression*), students who state that they have depression or an anxiety disorder (*diagnosed-all*), and students who state that they have never been diagnosed with any mental health conditions (*control*). We exclude 17 students who state they have a diagnosis other than depression or anxiety (e.g., an eating disorder) or who answered "don't know" or "prefer not to answer." Statistics of the dataset are displayed in Table 3. The dates of tweets in the dataset range from 2009 to 2020.

Table 4 reports the demographics of the users in the SURVEY-based dataset. The initial dataset was collected with the intention of providing a relatively balanced number of male and female students, and a similar number of students across class years. We also attempted to collect data from students from a variety of academic disciplines, but the imbalance across disciplines at the university made it impossible to balance these categories perfectly. Overall, we found that a higher percentage of women and freshmen provided their Twitter handles. Additionally, there is some imbalance introduced in these demographic categories when it comes to diagnosis, with a higher percentage of female students diagnosed with depression, along with a higher percentage of upperclassmen. The initial dataset was collected without attempting to balance the number of students based on race. Overall, the proportions are fairly reflective of the population of the university as a whole.

A limitation of our study is that this Twitter dataset is our only source of out-of-domain data that is not collected based on self-report, and it is relatively small. We would have preferred to test on multiple such datasets, but due to privacy concerns such data is very difficult to procure, and it is usually excluded entirely from NLP research on mental health. However, this dataset is especially suitable for studying generalization because it includes users who do not talk about mental health on social media. A second limitation, in addition to the size of our dataset, is that it represents the student population at a single university. This means that our results may only be representative of predictive power on this population, not on the population of Twitter overall. Collecting such a dataset that is perfectly representative of the Twitter population is not feasible, but our dataset may have specific biases, e.g., including mostly more educated users or wealthier users. Therefore, to complement our analysis of performance on out-of-domain data, we perform post-hoc analyses, including an analysis of feature weights in our classifier

[3]Community description: "For relatable posts that are too real for /r/meirl or /r/me_irl. Meaning jokes/posts about mental health issues and self deprecating humour."

[4]https://en.wikipedia.org/w/index.php?title=List_of_antidepressants&direction=next&oldid=1040008289.

[5]The data was collected as part of a study that underwent a full board review and was approved by the IRB at the University of Michigan (study number HUM0012629). All participants in the study have signed an informed consent form. 737 students completed the surveys, but we only include students who chose to provide active, public Twitter handles. The students were given $50 worth of gift cards for completion of 4 surveys.

[6]https://www.tweepy.org/.

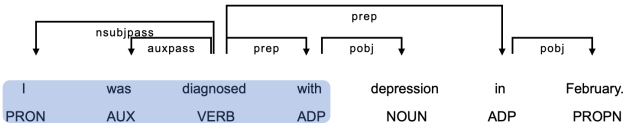

Figure 2: An example of dependency parsing tree of self-reported diagnosis post. Diagnosis pattern is highlighted in blue . In this example, we can extract diagnosis timestamp by following the path "diagnosed", "in", "February". We set the diagnosed date to February 14th, so we know that the potential error is $\pm$ two weeks.

that does not rely on our SURVEY-based dataset.

4 Diagnosis Timestamp Extraction

To study the temporal effect of diagnosis on depression classification methods, we extract the diagnosis timestamp for diagnosed users in the SELFREPORT dataset when possible. In the post in which users self-report their diagnosis, some users also share their diagnosis time.[7] For such users, we extract their diagnosis timestamp with two-week precision or better from the text of their self-report post.[8] To do this, we look only at the sentence where the self-reported diagnosis appears. From the sentence, we extract all time expressions that describe a DATE or TIME using SUTime (Chang and Manning 2012). However, extracted time expressions do not necessarily describe the diagnosis time. We further get the dependency parsing tree of the sentence using spaCy (Honnibal et al. 2020), and we only include time expressions that can be reached from the self-reported diagnosis pattern in the tree.[9] If no time expression can be reached, we go to the parent of the diagnosis pattern and check again. Finally, we remove time expressions that are unlikely to be precise within a two-week period such as "several months ago" and "years ago." Figure 2 shows an example where we can extract the precise timestamp "February." An example for which we cannot extract a precise diagnosis time is "Anyway, in 2017, I had my depression diagnosis," because "2017" is not precise on two-week level.

We are able to extract the diagnosis timestamp for 691 users (3.36% of total users) with two-week precision, and find that on average, there are 119.2 days between a user's self-report and their diagnosis. To evaluate the accuracy of the extracted diagnosis times, we randomly sample 100 users from the training set and manually annotate their diagnosis time. Our method achieves 96.3% precision on sampled data while retaining 59.1% recall.[10] Since our goal is to extract a precise diagnosis timestamp for users, we emphasize precision over recall.

[7] e.g., "I was diagnosed 3 months ago."

[8] If we switch to a slightly longer period such as one month, the number of users is similar. If we switched to a much longer period (one year window), we would have far more users ($> 3,000$), but this significant increase in diagnosis timeframe would introduce far more noise.

[9] When checking if a time expression can be reached, we exclude dependency relations ccomp, parataxis, conj, and advcl because we only want expressions that are related to diagnosis.

[10] Recall is measured by the percentage of users we extract a diagnosis time for among those who report a diagnosis time.

5 Experimental Setup and Results

To evaluate the effectiveness of using pre-diagnosis posts to improve models' generalizability, we train classifiers using the SELFREPORT dataset from different time periods (§5.1) and test their performance across those periods in an in-domain setting (§5.2). Then, we evaluate their ability to generalize to the SURVEY dataset and find that classifiers trained only on posts before diagnoses often outperform classifiers trained on all data in the transfer setting (§5.2). Finally, we further analyze the models to better understand their performance (§5.3).

5.1 Models

We consider four models ranging from Logistic Regression to Transformer-based language models (Vaswani et al. 2017; Ji et al. 2021). These models include larger language models, which typically take longer to train and are more difficult to interpret, but have state-of-the-art predictive power as well as smaller linear models, which may lead to lower accuracy, but are easier for stakeholders (e.g., mental health professionals) to understand. We search hyperparameters for each model on the validation set. Refer to the Appendix for details of the hyperparameter search and feature selection.

- **Logistic Regression**: we train two logistic regression models. The first one uses unigram and bigram **TF-IDF** features of the concatenated posts. The second one leverages Linguistic Inquiry and Word Count (**LIWC**) percentages (Pennebaker et al. 2015) of each post and uses their aggregation statistics (mean, variance, range, and quantile range) as features. In initial experiments, we used only the mean of LIWC values (as is common in NLP applications), but found that adding other aggregation statistics improved the performance on both SELFREPORT and SURVEY.
- **FastText** (Joulin et al. 2016): we train a FastText classification model using unigram and bigram features. FastText is an efficient implementation of linear classifiers on trained word embeddings. It achieves comparable or better performance with neural models on multiple mental health condition prediction tasks (Cohan et al. 2018). We concatenate all of a user's posts as the input.
- **MentalBERT** (Ji et al. 2021): we utilize the contextual representations generated by a BERT-like (Devlin et al. 2019) language model that is adapted to mental health related content from Reddit (Ji et al. 2021). **MentalBERT** is trained on seven mental health related subreddits, so it does not overlap with SURVEY. Among the seven subreddits, two of them could contain posts that overlap with data in SELFREPORT. We feed the BERT representation of each post to a feed forward neural network and aggregate by max pooling to get representations for each user.

We train each model on three subsets of data from SELFREPORT with different numbers of users and posts. We exclude mental health related data as described in §3.1.

- `All-large` contains posts from all time periods from all users, which is the common setting in previous work (Yates, Cohan, and Goharian 2017; Cohan et al. 2018).
- `Pre-diagnosis` considers diagnosed users for whom we can extract the diagnosis timestamp and their corresponding control users. Users also need to have at least one post before their diagnosis. We only keep posts before diagnosis for diagnosed users and randomly sample the same percent of posts for control users.
- `All-small` contains the same set of users as `Pre-diagnosis`, but it includes their posts from all time periods. This setting captures the impact of the time range of data, instead of the different set of users.

5.2 Results

In-domain performance. We first test classifiers on the SELFREPORT test set, using the same definitions of `All-large`, `All-small`, and `Pre-diagnosis`. Table 5 shows the results. We observe that classifiers that are trained on all data and all users (`All-large`) achieve the best performance in all test cases including the pre-diagnosis cases, demonstrating their good fit for the population of users who self-report their depression diagnoses on Reddit. However, their performance drops significantly when testing on only pre-diagnosis posts, indicating they focus more on features that appear after diagnosis. We find that traditional methods such as TF-IDF are surprisingly competitive with more recent models including FastText and MentalBERT. One reason for this might be that the data contains a number of "surface-level" features (e.g., n-grams) that make prediction relatively easy based on word counts. The TF-IDF model more directly represents these features, which may help the performance, especially on in-domain data.

Out-of-domain performance. Next, we evaluate classifiers on the SURVEY dataset. SURVEY represents a broader population in that it does not use self-reported diagnoses to identify diagnosed users; this means that users are less likely to explicitly mention their mental health. We only use it for testing purposes, and we include all tweets for each user. The results are presented in Table 6. We first notice that `Pre-diagnosis` classifiers achieve the best performance for four out of eight models; notably, when training on the same set and number of users (`All-small` vs. `Pre-diagnosis`), `Pre-diagnosis` data yields better or comparable performance on six of the eight models. This supports our hypothesis that training on pre-diagnosis posts generalizes better to the broader population. Training on `All-small` sometimes unexpectedly improves upon `All-large`; it is possible that there are unobserved demographic overlaps between `All-small` and SURVEY (a relatively homogenous group at one university) that cause this to occur. Given that, we focus primarily on the direct comparison between `All-small` and `Pre-diagnosis`.

For a broader set of mental health conditions (*diagnosed-all*, which includes users who have diagnosed anxiety in addition to users with diagnosed depression), training on pre-diagnosis posts provides stronger generalizability, as shown by larger gaps between `Pre-diagnosis` and `All-small` classifiers. We believe this generalization comes from some common symptoms shared by those with anxiety and depression (Hanson 2019). We were surprised to see a pattern of higher scores on the *diagnosed-all* dataset than the *diagnosed-depression* dataset, given that the users in the *diagnosed-depression* dataset all share a diagnosis with the users in the SELFREPORT dataset. From Table 3, we note that users in the *diagnosed-all* dataset tend to have more tweets, which could increase the likelihood that at least some of their tweets contain signals that are indicative of their symptoms.

Finally, we acknowledge that some of our best results come from using MentalBERT and FastText models with all data (`All-large`). With more computational resources, the MentalBERT model is presumably able to extract generalizable feature representations from the full dataset.[11] However, with smaller more interpretable models, which can be valuable in a healthcare settings, we find that the pre-diagnosis models generalize better.

5.3 Post-Hoc Analysis

Regression feature analysis. Table 7 illustrates the focus of `Pre-diagnosis` regression models on features that are more likely to occur post-diagnosis. For example, `All-large` focuses on n-grams such as "meds" and "disorder" whereas `Pre-diagnosis` captures "insecure" and "my life"; similarly, the LIWC classifier trained on `All-large` focuses on health- and anxiety-related words whereas `Pre-diagnosis` captures self-attentional focus, as indicated by first person singular pronouns (self preoccupation is known to relate with people's psychological status (Pennebaker 2004)). Although these features are indicative of depression for this specific population, as we show in §5.2, these post-diagnosis features do not always generalize to the broader population. Methods such as filtering lists of words that are directly related to mental health (as we do in §3.1) help to reduce reliance on these n-grams, but crafting these lists requires significant manual effort and subjective decisions. Changing the time period of data used reduces subjectivity and helps to filter out these features.

Effects of explicit mentions of mental health keywords. The features of the logistic regression classifiers lead us to believe that the `All-large` classifiers are likely to achieve better performance when users explicitly discuss mental health. We seek to quantify the extent to which that is the case by examining the correlation between the probability assigned to the depressed class for a user in the SURVEY dataset (averaged across five random seeds) and the frequency with which they use terms related to mental health in their tweets. Using the list of terms to identify Reddit posts related to mental health (see §3.1), we compute the percentage of words related to mental health in each of the depressed user's tweets.[12] Note that although we remove Reddit posts with these terms from the training data, these tweets

[11] FastText also took more than 10x the amount of time to train compared to TF-IDF and LIWC.

[12] For brevity, we only include users in Diagnosed-depression; the same patterns were present for users in Diagnosed-all.

Model \ Test Data		All-large	All-small	Pre-diagnosis
Random		18.18	18.18	18.18
TF-IDF	All-large	$\mathbf{72.05} \pm 1.36$	$\mathbf{72.07} \pm 1.36$	$\mathbf{64.05} \pm 1.68$
	All-small	69.32 ± 0.01	71.17 ± 0.00	61.30 ± 0.32
	Pre-diagnosis	60.50 ± 0.17	59.53 ± 0.27	62.35 ± 0.98
LIWC	All-large	$\mathbf{51.84} \pm 0.00$	$\mathbf{49.88} \pm 0.00$	$\mathbf{48.64} \pm 1.30$
	All-small	44.84 ± 0.00	44.03 ± 0.00	44.41 ± 0.52
	Pre-diagnosis	37.41 ± 0.10	36.39 ± 0.29	39.82 ± 0.51
FastText	All-large	$\mathbf{67.55} \pm 0.25$	$\mathbf{65.33} \pm 0.90$	$\mathbf{56.06} \pm 1.93$
	All-small	53.73 ± 0.44	52.97 ± 0.90	48.54 ± 1.92
	Pre-diagnosis	52.37 ± 0.52	52.06 ± 1.49	50.00 ± 1.54
MentalBERT	All-large	$\mathbf{75.01} \pm 0.80$	$\mathbf{74.52} \pm 0.21$	$\mathbf{63.32} \pm 1.77$
	All-small	68.94 ± 1.36	72.18 ± 1.63	62.13 ± 1.38
	Pre-diagnosis	59.48 ± 4.25	60.81 ± 5.08	60.09 ± 3.50

Table 5: F1 score for diagnosed users on SELFREPORT (mean across five random seeds; error shows standard deviation). The best performance in each cell is in bold. All-large classifiers achieve the best performance in most cases, but their performance drops significantly when testing only on pre-diagnosis posts.

		TF-IDF	LIWC	FastText	MentalBERT
Diagnosed-depression	Random		27.35		
	All-large	42.76 ± 2.41	40.82 ± 0.00	35.70 ± 8.88	$\mathbf{45.70} \pm 2.67$
	All-small	41.18 ± 0.00	$\mathbf{40.86} \pm 0.00$	$\mathbf{39.22} \pm 2.06$	43.52 ± 2.80
	Pre-diagnosis	$\mathbf{43.40}^{*} \pm 0.00$	40.36 ± 1.10	34.84 ± 3.96	41.19 ± 1.73
Diagnosed-all	Random		36.30		
	All-large	44.45 ± 2.49	43.41 ± 0.00	30.64 ± 10.21	$\mathbf{45.63} \pm 2.26$
	All-small	47.06 ± 0.00	46.03 ± 0.00	40.29 ± 0.41	44.21 ± 1.55
	Pre-diagnosis	$\mathbf{48.57}^{*} \pm 0.00$	$\mathbf{48.65}^{*} \pm 0.64$	$\mathbf{40.39} \pm 3.69$	45.17 ± 2.18

Table 6: F1 score for diagnosed users on SURVEY (mean across five random seeds; error shows standard deviation). The best mean results are in bold. * indicates statistically significant improvement of Pre-diagnosis from All-small (Wilcoxon signed-rank test; Wilcoxon 1945, $p \leq 0.05$).

are not removed from the test data as we intend for our test dataset to represent all text written by users who were not identified by mental-health related content.

We find that the Spearman's rank correlation coefficient between the percentage of tokens related to mental health and $P(depressed)$ predicted by the classifier was positive and statistically significant ($p < 0.01$) for all classifiers (Table 8). In the case of TF-IDF, FastText, and MentalBERT, the correlations drop between All-small and Pre-diagnosis, indicating that the classifiers rely less on features corresponding to explicit mentions of mental health.[13] The only classifier for which the correlation increases was based on LIWC features; this may be because terms are mapped to 73 discrete LIWC categories, and a large number of terms are not present in those categories.

[13] Although these terms are excluded from the training set, their presence likely is correlated with other terms that are not excluded, such as those in Table 7.

Varied time periods using in-domain data. Next, we explore how the performance changes when we vary the time period covered by the test data. Concretely, we evaluate classifiers on diagnosed users using their posts before a specific time point. To exclude the influence of other factors, we test on the same set of diagnosed users who have at least one post 90 days before their diagnoses, and we down-sample users' posts to keep the number of posts unchanged for different test time points. We also include the corresponding control users in the test set and down-sample their posts proportionally to the number of posts of diagnosed users. As shown in Figure 3, Pre-diagnosis has a relatively consistent performance for different time periods. It achieves better or comparable performance compared to All-small classifiers, excluding LIWC where the performance is consistently worse. It is also worth noting that the performance of All-large classifiers keeps increasing as we include more posts after diagnoses in the test set, indicating All-large classifiers focus more on signals that are prevalent after diagnoses.

	Top depression features
TF-IDF `All-large`	mental health, meds, medication, mental, anxious, lonely, kill myself, myself, disorder, my doctor
TF-IDF `Pre-diagnosis`	insecure, find his, someone better, my life, my dad, okay, parents, friends, told, are these
LIWC `All-large`	*HEALTH*, *ANX*, CONJ, PREP, NEGEMO,ANX, CONJ, BIO, NUMBER, I
LIWC `Pre-diagnosis`	I, *HEALTH*, I, *I*, NEGEMO, BIO, FUNCTION, INSIGHT, INTERROG, *HEAR*

Table 7: Top 10 positive features for TF-IDF and LIWC classifiers ordered by weight. For LIWC, text styles are used to represent different aggregation measures across user's posts: mean, 75 percent range, and *90 percent range*. `All-large` classifier focuses on indicative features that appear more after diagnoses (e.g., "medication" and *ANX*). `Pre-diagnosis` classifier captures more robust features such as self preoccupation (e.g., "my life" and I).

	TF-IDF	LIWC	FastText	MentalBERT
All-Large	0.87	0.58	0.69	0.87
All-Small	0.86	0.52	0.57	0.79
Pre-diagnosis	0.82	0.61	0.50	0.74

Table 8: Spearman's ρ coefficient between probability of depression predicted by each classifier for students in the Diagnosed-depression group and the percentage of tokens in our mental health keyword list. $p < 0.01$ for all models.

Finally, we fit a line for each of the plots in Figure 3 and report the slopes in Table 9. `All-large` and `All-small` classifiers have the largest slope in all cases, and `Pre-diagnosis` classifiers have the smallest slope in most cases, confirming that classifiers that train on all posts focus on post-diagnosis features whereas `Pre-diagnosis` classifiers are more consistent.

	TF-IDF	LIWC	FastText	MentalBERT
All-large	**9.07**	10.08	3.65	**21.57**
All-small	-5.05	**12.11**	**14.89**	18.29
Pre-diagnosis	0.82	9.80	3.24	11.90

Table 9: Slope of the fitted lines in Figure 3 (on the order of 10^{-5}). The largest value for each model is in bold, and reflects the focus on post-diagnosis features.

Experiments on additional data. Our Twitter survey-based dataset is ideal for our use case because it represents the social media activity of depressed users who are not identified using self-reports, however the size of the dataset is limited. We are not aware of similar datasets that are openly available, but we test our classifiers on a related dataset, the CLPsych 2015 Shared Task dataset (Coppersmith et al. 2015) which was collected using *self-reports* on Twitter. The test split of the dataset includes 150 depressed users and 300 control users (we exclude users with PTSD).

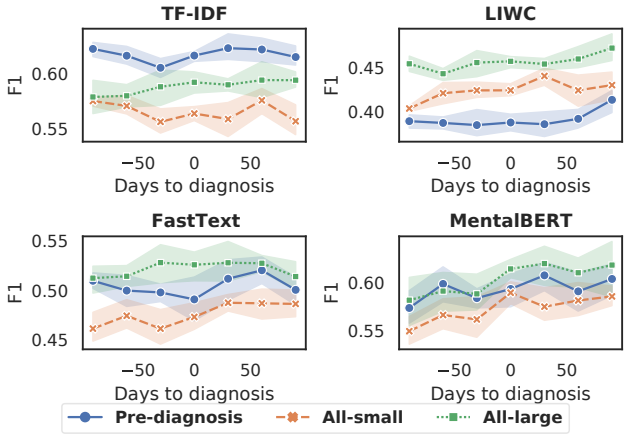

Figure 3: F1 score for diagnosed users tested on SELFREPORT (average of five runs). Shaded area shows the 95% confidence interval. For diagnosed users, we only consider posts before a certain time point (x axis), and we downsample the posts so that each user has the same number of posts at each test point.

We report the classification results using classifiers trained on our Reddit dataset and tested on the CLPsych test data in Table 10. We find that `Pre-diagnosis` achieved comparable performance to `All-small` using the TF-IDF, LIWC, and FastText models. When these results are compared to the results on the `All-large` test set in Table 5, we find that the difference between the two models decreases when considering out-of-domain test data.

6 Discussion

In our experiments, we find that classifiers using exclusively `Pre-diagnosis` data often outperform classifiers using all data from all users with self-reported depression when tested on a population that does not self-report their diagnosis online. This trend is slightly amplified when we compare `Pre-diagnosis` classifiers with classifiers trained on the same set of users, but with all data (`All-small`). While we do not see the same trend on in-domain data, we find that when posts are downsampled such that classifiers are tested on the same number of posts per user, the `Pre-diagnosis` model achieves comparable results to others that are trained on far more data. By focusing on pre-diagnosis data, we also demonstrate an approach that would likely generalize particularly well to people who are not yet diagnosed and need to be connected to help (Figure 1).

This is important in that (a) it shows that a more focused approach in the data curation stage can help with the generalization issues noted by Harrigian, Aguirre, and Dredze (2020) and Ernala et al. (2019), and (b) we can achieve

	TF-IDF	LIWC	FastText	MentalBERT
Random		40.00		
`All-large`	**63.24**±0.73	**55.74**±0.00	44.34±4.45	**67.70**±3.21
`All-small`	58.53±0.06	52.86±0.00	51.78±2.24	66.06±1.49
`Pre-diagnosis`	57.82±0.14	**55.74**±0.30	**52.62**±0.88	59.45±1.58

Table 10: F1 score for depressed users in CLPsych (mean across five random seeds; error shows standard deviation). The best performance in each column is in bold.

strong results with far less data, reducing the necessary computational resources (for the MentalBERT model, training on `Pre-diagnosis` takes less than 1% of the training time of the `All-large` model on the same device). Our study shows that relying exclusively on **big data** is not enough to build effective classifiers; rather, **data quality** is critical. Improving data quality is possible by re-examining assumptions in the data-curation process. Furthermore, the results shown in Tables 5 and 6 demonstrate that the performance differences that are salient on in-domain data are not necessarily representative of what we may see when using out-of-domain data. While classifiers using All-Large data outperform those that use All-Small or Pre-Diagnosis on in-domain test data in every setting we examined (across four classifiers and three test data subsets), the same is not true with out-of-domain data. We believe that Pre-Diagnosis data should be considered when attempting to identify users in need of support who we do not expect to openly discuss depression, especially when an interpretable model is desired. The feature weights in Table 7 along with the correlations in Table 8 support this conclusion, as they show the ability of Pre-Diagnosis classifiers to move past classifying based primarily on discussion of mental health.

To the best of our knowledge, this is the first study to investigate how the time period from which training data is extracted (with respect to diagnosis) affects the generalization of mental health classifiers to different time periods and domains. Table 7 shows that the features associated with depression differ when we use pre-diagnosis data. A similar phenomena has been shown in prior work comparing data from various types of subreddits and even data from non-clinical subreddits before vs. after the first clinical subreddit post (Thorstad and Wolff 2019), but we are the first to explicitly identify a diagnosis date from self-report posts. We use that date not only to demonstrate a difference in the features that are present, but also to show that when trained on the same set of users, we improve generalization to out-of-domain data. Our findings open up the possibility of exploring the same phenomena on other mental health diagnoses (e.g. anxiety). Additionally, they reinforce the need to consider temporal variation in classification problems that aim to classify people based on text they write over time but considers their behavior to be static.

This study also opens up an opportunity to consider ways to attach meaningful temporal annotations to data such as diagnosis date or the date at which a user begins therapy. Although the fact that the `Pre-diagnosis` dataset is substantially smaller than `All-large` could help in low-compute settings, it could also be viewed as a limitation. We chose a pattern-matching method with high precision to identify diagnosis dates, but future work could use modeling approaches to determine whether identifying noisier diagnosis dates for more users leads to improved models.

7 Conclusion

In this paper, we proposed and validated the hypothesis that using only data from before users are diagnosed can lead to more robust features for mental health classifiers. In particular, we focused on addressing the generalization problems that arise from datasets built on self-reported diagnoses. As the users in such datasets by definition talk about mental health online, classifiers built on these datasets often rely on overt mentions of mental health and symptoms, which are not as prevalent in datasets not built on self-reports. We showed that these features are less prevalent before users are diagnosed, and thus using pre-diagnosis data from a self-report-based dataset helps to avoid this bias, and sometimes improves generalization. Our results showed that reconsidering the set of data that we use for training can lead to improved performance, especially with smaller, more interpretable models that focus on human behavior. In the future, we believe that precise data-curation methods can be used in conjunction with modeling techniques that aim to improve generalizability (Lee et al. 2021; Nguyen et al. 2022) to build more robust classifiers.

We believe our study can have implications on current datasets and models for mental health classification, by reconsidering what portion of the data is being used for model training. The possibility of using a smaller training dataset for comparable results on out-of-domain data will also increase the applicability of these models, by allowing to be deployed in low compute settings. Additionally, while the current medical approach to mental health is mostly focused on the treatment of an already existing mental health condition, we believe our study can open the doors to more preventative approaches, where people who are more likely to experience mental health issues in the future are identified early to prevent the progression of the illness.

A Classification Models

We provide more details about our classification models here, including selected hyperparameters and feature selection details. All hyperparameters are searched on the corresponding validation set.

Hyperparameter	Value
regularization strength searched	[0.3, 1, 3]
TF-IDF min_df	5
TF-IDF max_df	$0.7 * \|D\|$
TF-IDF max # features	100000

Table 11: Hyperparameters used to train TF-IDF and LIWC models. max_df and min_df mean the maximum and minimum document frequency. $|D|$ is the number of documents in the training set.

A.1 Logistic Regression Models

We use implementation from scikit-learn for the logistic regression model and TF-IDF vectorizer (Pedregosa et al. 2011). For the **TF-IDF** classifier, we use the combination of unigram and bigram features. For the **LIWC** classifier, we use five aggregation statistics on user posts: mean, variance, range, 90 percent range, and 75 percent range. The hyperparameters are listed in Table 11. All hyperparameters that are not listed take the default values in scikit-learn.

A.2 FastText Models

We use the implementation[14] in Joulin et al. (2016). We use unigram and bigram features and a minimum word occurrence of 10. Our hyperparameter search includes learning rates [0.4, 0.5, 0.6, 0.7] and # of epochs [40, 50, 60, 70].

A.3 MentalBERT Models

We use the MentalBERT model to generate representations for each user post (Ji et al. 2021). The model is trained on user posts from seven mental-health related subreddits: r/depression, r/SuicideWatch, r/Anxiety, r/offmychest, r/bipolar, r/mentalillness/, and r/mentalhealth. Among these subreddits, r/offmychest and r/mentalillness/ might contain posts that overlap with data in SELFREPORT. For efficiency, we use the fixed representations (`[CLS]` token) generated by MentalBERT without fine-tuning. We pass the MentalBERT representations to a Feedforward Neural Network (FNN) and use max pooling to get aggregated user representation.

We train all models on a GeForce RTX 2080 Ti GPU. Generating the fixed MentalBERT representations takes around 29 hours. The training time is 8 hours for the `All-large` model and 3 minutes for the `All-small` and `Pre-diagnosis` models. The hyperparameters for the **MentalBERT** classifiers are in Table 12.

Hyperparameter	Value
number of epochs	20
patience	3
maximum learning rate searched	[0.0001, 0.001]
learning rate scheduler	linear decay
optimizer	Adam (Kingma and Ba 2017)
weight decay searched	[0.0005, 0.005]
Adam beta weights	0.9, 0.999
dimension of FNN linear layer 1 linear layer 2 linear layer 3	 512 128 128
dropout in FNN	0.1

Table 12: Hyperparameters used to train MentalBERT models.

Ethical Statement

While the users in our training set all shared their depression diagnosis on a public online forum, we acknowledge that special care should be taken with such data considering the sensitivity of the subject. As has been done in prior studies, we removed identifiers such as Reddit usernames from our personal copy of the data, and make no attempt to ascertain any information about the users who comprise our dataset beyond what is written in their Reddit posts. The study that resulted in the Twitter dataset received full IRB approval from our institution; personal identifiers such as Twitter usernames were scrubbed from the dataset. Ethics around health-related social media data are explored in more detail in Benton, Coppersmith, and Dredze (2017).

In our work, we show one method that improves generalizability of depression diagnosis classifiers (with simpler models) to a population of people who may not explicitly discuss mental health. While our method *improves* upon the baseline, the results **do not suggest that such a model is ready for real-world deployment**. The task of detecting depression from text is very challenging, and our results on out-of-domain data that is collected without using self-reported diagnoses show that we still have a long way to go with respect to accuracy on populations that differ from those that we see in our training data (regardless of how that data is sampled).

However, the more important question to ask may be how such a classification system should be used *if accuracy reaches an acceptable threshold* and *how to define that threshold for various potential applications*. A very accurate depression classification system could be used for good: monitoring population-level depression (e.g., Wolohan (2020)), routing counselors to those with the most need in resource-constrained settings (e.g., as recommended by Bantilan et al. (2020)), opt-in prompts to receive counseling on college campuses if classifiers see symptoms developing, or opt-in monitoring for people who are already receiving counseling. However, the same systems could also be used for nefarious purposes, such as denying jobs to people whose mental health status is inferred from their social media posts. This would be illegal in the United States, but it may not be in all countries, and an action being illegal

[14] https://fasttext.cc

does not eliminate the risk of it occurring. While out-of-scope for this paper, the question of how mental health classifiers should be used and which classification setups will most benefit society while reducing harm should be considered more thoroughly by the community, with active involvement from mental health practitioners. To the best of our knowledge, these considerations have been understudied in the NLP community; the few exceptions that focus on the ethical tensions surrounding mental health classifiers have appeared outside of NLP (Chancellor et al. 2019). We hope that the community will consider and participate in inner-disciplinary work that directly considers how mental health classification models can be deployed; one recent example of such work is Cohen et al. (2020).

Acknowledgements

This work was partially supported by the University of Michigan Eisenberg Family Depression Center and by the Templeton Foundation (#62256).

References

Bantilan, N.; Malgaroli, M.; Ray, B.; and Hull, T. D. 2020. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. *Psychotherapy Research*, 31(3): 289–299.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.

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. Valencia, Spain: Association for Computational Linguistics.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research*, 19(8): e289.

Campillo-Ageitos, E.; Martinez-Romo, J.; and Araujo, L. 2022. UNED-MED at eRisk 2022: depression detection with TF-IDF, linguistic features and Embeddings. *Working Notes of CLEF*, 5–8.

Chancellor, S.; Birnbaum, M. L.; Caine, E. D.; Silenzio, V. M. B.; and De Choudhury, M. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 79–88. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Chang, A. X.; and Manning, C. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3735–3740. Istanbul, Turkey: European Language Resources Association (ELRA).

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

Cohen, J.; Wright-Berryman, J.; Rohlfs, L.; Wright, D.; Campbell, M.; Gingrich, D.; Santel, D.; and Pestian, J. 2020. A Feasibility Study Using a Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in Adolescent Therapy Sessions. *International Journal of Environmental Research and Public Health*, 17(21): 8187.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60. Baltimore, Maryland, USA: Association for Computational Linguistics.

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

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression via Social Media. In *International AAAI Conference on Web and Social Media*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D. A.; and Schwartz, H. A. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44): 11203–11208.

Ernala, S. K.; Birnbaum, M. L.; Candan, K. A.; Rizvi, A. F.; Sterling, W. A.; Kane, J. M.; and De Choudhury, M. 2019. Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Gkotsis, G.; Oellrich, A.; Velupillai, S.; Liakata, M.; Hubbard, T. J. P.; Dobson, R. J. B.; and Dutta, R. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports*, 7: 45141.

Hanson, J. 2019. Identifying anxiety, depression signs. https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/addressing-your-mental-health-by-identifying-the-signs-of-anxiety-and-depression. Accessed: 2023-04-21.

Harrigian, K.; Aguirre, C.; and Dredze, M. 2020. Do Models of Mental Health Based on Social Media Data Generalize?

In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3774–3788. Online: Association for Computational Linguistics.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://spacy.io/. Accessed: 2023-04-21.

Jamil, Z.; Inkpen, D.; Buddhitha, P.; and White, K. 2017. Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 32–40. Vancouver, BC: Association for Computational Linguistics.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. *arXiv preprint arXiv:2110.15621*.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lee, A.; Kummerfeld, J. K.; An, L.; and Mihalcea, R. 2021. Micromodels for Efficient, Explainable, and Reusable Systems: A Case Study on Mental Health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4257–4272. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Losada, D. E.; Crestani, F.; and Parapar, J. 2018. Overview of eRisk: Early Risk Prediction on the Internet. In Bellot, P.; Trabelsi, C.; Mothe, J.; Murtagh, F.; Nie, J. Y.; Soulier, L.; SanJuan, E.; Cappellato, L.; and Ferro, N., eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 343–361. Cham: Springer International Publishing. ISBN 978-3-319-98932-7.

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.

Massanari, A. L. 2016. Contested play: The culture and politics of reddit bots. In *Socialbots and their friends*, 126–143. Routledge.

McManus, K.; Mallory, E. K.; Goldfeder, R. L.; Haynes, W. A.; and Tatum, J. D. 2015. Mining Twitter Data to Improve Detection of Schizophrenia. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2015: 122–126.

Mitchell, M.; Hollingshead, K.; and Coppersmith, G. 2015. Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 11–20. Denver, Colorado: Association for Computational Linguistics.

Mukherjee, S.; and Das, S. 2022. Application of transformer-based language models to detect hate speech in social media. *Journal of Computational and Cognitive Engineering*.

Nguyen, T.; Yates, A.; Zirikly, A.; Desmet, B.; and Cohan, A. 2022. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8446–8459. Dublin, Ireland: Association for Computational Linguistics.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pennebaker, J. W. 2004. Theories, Therapies, and Taxpayers: On the Complexities of the Expressive Writing Paradigm. *Clinical Psychology: Science and Practice*, 11(2): 138 – 142.

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

Shen, J. H.; and Rudzicz, F. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 58–65. Vancouver, BC: Association for Computational Linguistics.

Thorstad, R.; and Wolff, P. 2019. Predicting future mental illness from social media: A big-data approach. *Behavior Research Methods*, 51(4): 1586–1600.

Uban, A.-S.; Chulvi, B.; and Rosso, P. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124: 480–494.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83.

Wolohan, J. 2020. Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics.

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 *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 11–21. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Yates, A.; Cohan, A.; and Goharian, N. 2017. Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2968–2978. Copenhagen, Denmark: Association for Computational Linguistics.