# Detection of Mental Health Conditions from Reddit via Deep Contextualized Representations

**Zhengping Jiang**
Computer Science Dept.
Columbia University
zj2265@columbia.edu

**Sarah Ita Levitan**
Computer Science Dept.
Hunter College, CUNY
sarah.levitan@hunter.cuny.edu

**Jonathan Zomick**
Psychology Dept.
Hofstra University
jzomick1@pride.hofstra.edu

**Julia Hirschberg**
Computer Science Dept.
Columbia University
julia@cs.columbia.edu

## Abstract

We address the problem of automatic detection of psychiatric disorders from the linguistic content of social media posts. We build a large scale dataset of Reddit posts from users with eight disorders and a control user group. We extract and analyze linguistic characteristics of posts and identify differences between diagnostic groups. We build strong classification models based on deep contextualized word representations and show that they outperform previously applied statistical models with simple linguistic features by large margins. We compare user-level and post-level classification performance, as well as an ensembled multiclass model.

## 1 Introduction

Global prevalence of mental disorders has been estimated at 29.2% in a meta-study of 174 surveys across 63 countries (Steel et al., 2014). Mental illness is one of the leading causes of disability globally and the costs of mental health treatment have run into the trillions of dollars (Organization et al., 2014; Vigo et al., 2016; Patel et al., 2018). Additionally, individuals suffering from mental illness are estimated at forming 14.3% of deaths worldwide, significantly higher than a control population (Walker et al., 2015). Limited mental health resources and funding have necessitated new approaches to addressing the global impact of this problem. However, early detection of mental illness and early intervention have shown promising results for improving treatment and long-term outcome results for many psychiatric disorders; these have the potential to reduce the costly burden that mental illness has placed on our society as well as our global economies (Bird et al., 2010; Treasure and Russell, 2011; De Girolamo et al., 2012; Murru and Carpiniello, 2018).

Advances in artificial intelligence in general and computational linguistics in particular have made important contributions to detecting and predicting mental illness among the population, particularly in social media (Guntuku et al., 2017; Wongkoblap et al., 2017). Using computational linguistics, researchers have been able to leverage the widespread use of social media to analyze large, publicly available datasets for identifying linguistic markers of mental illness. To date, unique linguistic markers and patterns have been identified for several psychiatric conditions, such as major depressive disorder (MDD)(De Choudhury et al., 2013; Vedula and Parthasarathy, 2017), general anxiety disorder (GAD) (Shen and Rudzicz, 2017), bipolar disorder (BD) (Huang et al., 2017; Sekulić et al., 2018), eating disorders (ED) (Mohammadi et al., 2019; Naderi et al., 2019), schizophrenia (SZ) (Mitchell et al., 2015; Birnbaum et al., 2017; Zomick et al., 2019), obsessive compulsive disorder (OCD) (Coppersmith et al., 2015a), posttraumatic stress disorder (PTSD) (Coppersmith et al., 2014), as well as others (Coppersmith et al., 2015a). Linguistic findings have spanned various domains of language, including the use of pronouns, emotion words, tentative language, tangentiality, punctuation, and content analysis. The majority of these models have been developed to successfully predict if a given user has self-disclosed receiving a diagnosis for a psychiatric condition and is currently suffering with mental illness.

However, much of this previous research on social media and mental health has focused on comparing users with particular disorders with control users. In this work we expand this focus to compare across a wide set of common disorders. This is directly applicable to real-world diagnostic scenarios, where clinicians select a diagnosis from a large set of disorders, rather than simply diagnosing an individual as healthy or not. In addition, prior work



*Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156

https://doi.org/10.18653/v1/P17

has focused on data collection and analysis, with less emphasis on building strong predictive models. In this work, we apply state-of-the-art neural network models developed for other natural language tasks to the problem of mental health detection from social media.

## 2 Related Work

In recent years, there has been increased interest in the NLP community in the automatic detection of psychiatric conditions from language. Many researchers have focused on analyzing vast amounts of language from social media posts to study mental health (Birnbaum et al., 2017; Coppersmith et al., 2015a; Mitchell et al., 2015). With the advent of social media, many people who suffer from various forms of mental illness have found a sense of community and support, and these platforms offer a mode of expression for discussing their experiences openly online. Additionally, many online platforms allow users to post anonymously, giving them a sense of security and anonymity to discuss their experiences and struggles without the fear of being stigmatized or discriminated against (Balani and De Choudhury, 2015; Berry et al., 2017; Highton-Williamson et al., 2015).

In order to analyze language patterns related to various disorders from social media data, researchers have developed innovative approaches for automatically labeling this data. (Coppersmith et al., 2014) developed a widely used approach for gathering data for a range of psychological disorders, using regular expressions to identify public self-disclosures of diagnoses on social media. They tested this approach using Twitter data and collected a dataset of tweets from individuals with bipolar, depression, PTSD, SAD, and a control group. They analyzed several linguistic features across conditions using a clustering algorithm and built predictive classifiers to distinguish between diagnosed and control users. (Cohan et al., 2018a) expanded this approach to study a larger set of disorders using Reddit data. Reddit is one of the fastest growing and widely used social media platforms, averaging over 330 million active monthly users, and, as of 2018, was the fourth most visited website in the US (Hutchinson, 2018). Unlike Twitter, Reddit imposes no limits on the length of posts, enabling an analysis of longer language samples. In addition, Reddit is composed of subreddits, which are forums dedicated to specific topics, and there are many subreddits related to specific mental health conditions. They collected a large dataset of Reddit posts and analyzed linguistic features between different conditions and a control group. They also trained binary classifiers to distinguish between each condition and the control.

Our work directly builds on this prior work. Following (Coppersmith et al., 2014) and (Cohan et al., 2018a), we collect a large expanded dataset of Reddit posts. Unlike prior work, we do not focus on pairwise analyses of linguistic features between conditions and the control group; rather, we compare features between conditions to highlight important differences that can distinguish between various disorders. While others have trained simple predictive models of these disorders, we instead use state-of-the-art deep contextualized models that have been highly successful across several NLP tasks. Our work makes important contributions to the problem of mental health detection from social media data and provides insights for others to further build on this work.

## 3 Data Collection

We focus in this study on 8 mental health conditions: schizophrenia (SZ), borderline personality disorder (BPD), post-traumatic stress disorder (PTSD), eating disorder (ED), major depression disorder (MDD), general anxiety disorder (GAD) and bipolar disorder. While datasets for many of these conditions have been collected on varying scales, to the best of our knowledge our dataset includes the largest cohort of users whose posts have been collected for many of these conditions. To build this cohort, we collect users with self-identified mental health conditions from Reddit using the Pushshift API[1]. We search for users in mental health related subreddits and use keywords to search for mental health related words. Our distant labeling approach is further explained below, in Section 3.1. We also identify a group of control users who do not have any of the targeted conditions. We first collect a large scale user pool by scraping posts from common subreddits like r/AskReddit, and filter the control users by process described in subsection 3.2. The number of posts collected in each condition is shown in Table 1 and the number of users whose posts were collected in each is shown in Table 2.

[1] https://github.com/pushshift/api

| | Post Number | Avg. Token |
|---|---|---|
| SZ | 1084k | 43.7 |
| BPD | 1629k | 43.6 |
| PTSD | 2169k | 46.1 |
| ED | 396k | 43.0 |
| MDD | 1585k | 42.9 |
| GAD | 3047k | 42.2 |
| OCD | 1813k | 38.6 |
| Bipolar | 5819k | 40.5 |
| Total | 17.5m | 42.0 |

Table 1: Dataset Statistics (Posts)

| | User Num | Unique | Clf. |
|---|---|---|---|
| SZ | 2134 | 1741 | 1175 |
| BPD | 4695 | 3430 | 2275 |
| PTSD | 5294 | 3840 | 2666 |
| ED | 1005 | 752 | 514 |
| MDD | 3183 | 1832 | 1360 |
| GAD | 4958 | 3155 | 2388 |
| OCD | 4151 | 3140 | 2211 |
| Bipolar | 11186 | 9524 | 6420 |
| Total | 35606 | 27214 | 19009 |

Table 2: Dataset Statistics (Users), where **Unique** column figures correspond to number of users without comorbidity issue and **Clf.** column figures correspond to number of users we use for our classification task.

### 3.1 Distant Labeling

We generally follow the self-identification technique previously employed in (Mitchell et al., 2015; Coppersmith et al., 2015a; Cohan et al., 2018a). Specifically, we construct separate regular expressions for self-identification checking and condition resolution. We use 2-way human annotation to verify the performance of our labeling algorithm. Our second version of labeling algorithm achieves high precision (over .95) when tested on a held-out validation set. We found that posts directly identifying with "eating disorder" are scarce, so we collapse identification with "anorexia", "arfid", "bulimia" and "binge" into a single category for "eating disorder". We also calculate comorbidity statistics for our extracted user set as is shown in Figure 1, and have found it correlated well with statistics previously reported (Coppersmith et al., 2015a; Cohan et al., 2018b).

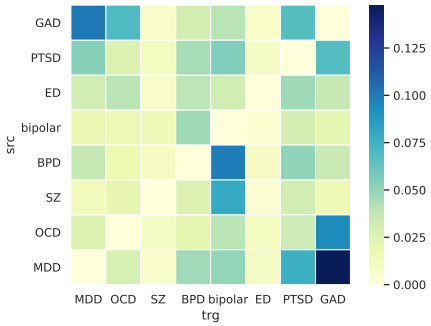


Figure 1: Comorbidity matrix of our dataset, each ceil corresponds to the portion of users of the src condition that have the trg condition.

### 3.2 Preprocessing

Following Cohan et al.(2018b) we do not include in our classification any control user that has any *sensitive post*, defined as either (1) containing mental health related keywords or (2) posted in a mental health related subreddit. In addition, under the (**CL**) condition of our classification experiments (described below in Section 6), we remove these sensitive posts from mental group users. For post level preprocessing, we replace emojis with descriptive text using the demoji package[2], normalize html characters like "", "&" and " " etc., and mask out url, email and subreddit references with regular expression.

## 4 Linguistic Indicators of Mental Health

After collecting and preprocessing the data, we analyzed linguistic characteristics of mental health using Linguistic Inquiry and Word Count (Pennebaker et al., 2015). LIWC is a text analysis program that computes word counts for semantic classes and structural features. It relies on an internal dictionary that maps words to psychologically motivated categories. These include standard linguistic features (e.g. percentage of words that are pronouns, articles), markers of psychological processes (e.g. affect, social, cognitive words), and punctuation categories (e.g. periods, commas). LIWC dimensions have been used in many studies to predict outcomes including personality (Pennebaker and King, 1999), deception (Newman et al., 2003), and health (Pennebaker et al., 1997). We extracted 73 features using LIWC 2015; a full description of these features is found

[2]https://pypi.org/project/demoji/

in (Pennebaker et al., 2015). To construct a single feature vector per user, we concatenated all posts per user and then extracted the LIWC features from the combined posts and performed length normalization.

Prior work on identifying linguistic indicators of mental health has compared LIWC features from users' individual disorders with healthy control users. However, it is often unclear whether the findings are specific to the disorder, or if they are indicative of mental disorders more generally. For example, in pairwise analyses, personal pronoun usage has been found to be increased in individuals with schizophrenia (Zomick et al., 2019). However, this pattern might or might not be specific to schizophrenia, but may be indicative of other mental disorders as well. Because of this gap in prior work, we began by comparing LIWC features directly across the 8 diagnostic groups and the control group.

Figure 2 shows a heatmap of the z-score normalized average LIWC features across users in each group. The x-axis shows the 8 diagnostic groups and the control group, and the y-axis shows the normalized LIWC feature values. The color of each cell indicates whether the scaled value is high (blue), low (red) or average (white). As shown in this figure, the control group has the greatest number of red features, or LIWC features which have a low frequency. It is clear from the figure that the 8 diagnostic groups have different language usage patterns from the control group, and particularly show a higher frequency for several linguistic dimensions. Further, there seem to be several interesting similarities and differences in linguistic patterns across the diagnostic groups. To further investigate these differences, we ran one-way ANOVAs comparing each LIWC feature across the 8 diagnostic groups and the control group. To correct for family-wise type I errors, we used Bonferroni correction. The results indicated that there were significant differences across groups for all 73 LIWC variables. We ran Tukey posthoc tests to identify which pairs of conditions were most similar and most different. Because of limited space, we focus here on the linguistic dimensions with the greatest variance among the groups, indicated by the highest F-statistics. These categories were *anx*, the use of anxiety words ($F(8, 24442) = 531.911$, $p<.0001$), and *I*, the use of the first person singular pronoun ($F(8, 24442) = 438.738$, $p<.0001$). Figure 3 shows the results of the posthoc analysis. Pairwise comparisons among psychiatric conditions revealed several interesting findings. While each condition differed significantly for both features when compared with the control group (users in the control group were significantly less likely to use anxiety related words and "I" when compared to each condition), when comparing between the psychiatric conditions differences varied. For example, users with SZ were significantly less likely to use anxiety related words in comparison with other groups. Another interesting finding was that users with BPD used 1st person singular pronouns significantly more than all other psychiatric conditions with the exception of ED. These findings shed light on linguistic variation across different psychiatric conditions, and provide further motivation for developing methods to distinguish between individuals with these disorders by leveraging social media posts.

## 5 Methods for Classification Experiments

Having identified significant differences in linguistic features between the disorders, and between the control users, we next explore several classification methods for automatically identifying different mental health conditions. Previous efforts to identify such conditions in Reddit have primarily employed simple logistic regression or SVMs using a bag-of-words representation or LIWC features, and some have explored RNN/CNN based text encoder models (Coppersmith et al., 2014, 2015a,b; Cohan et al., 2018b; Sekulic and Strube, 2019). However, recent advances in contextual representations like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which have enabled substantial performance increases across many NLP task, have not been well integrated into mental health identification tasks. This is due to model size and scalability issues of the large number of posts generated by a user. In this work we focus on methods utilizing contextual representations for mental health identification and compare their effectiveness to a logistic regression baseline trained on LIWC features. We present an attention-based model using BERT representations as input features, as well as a REALM-like model (Guu et al., 2020) inspired by recent advances in open domain question-answering. All of these models are trained for a user-level classification task, to detect whether a user has a particular

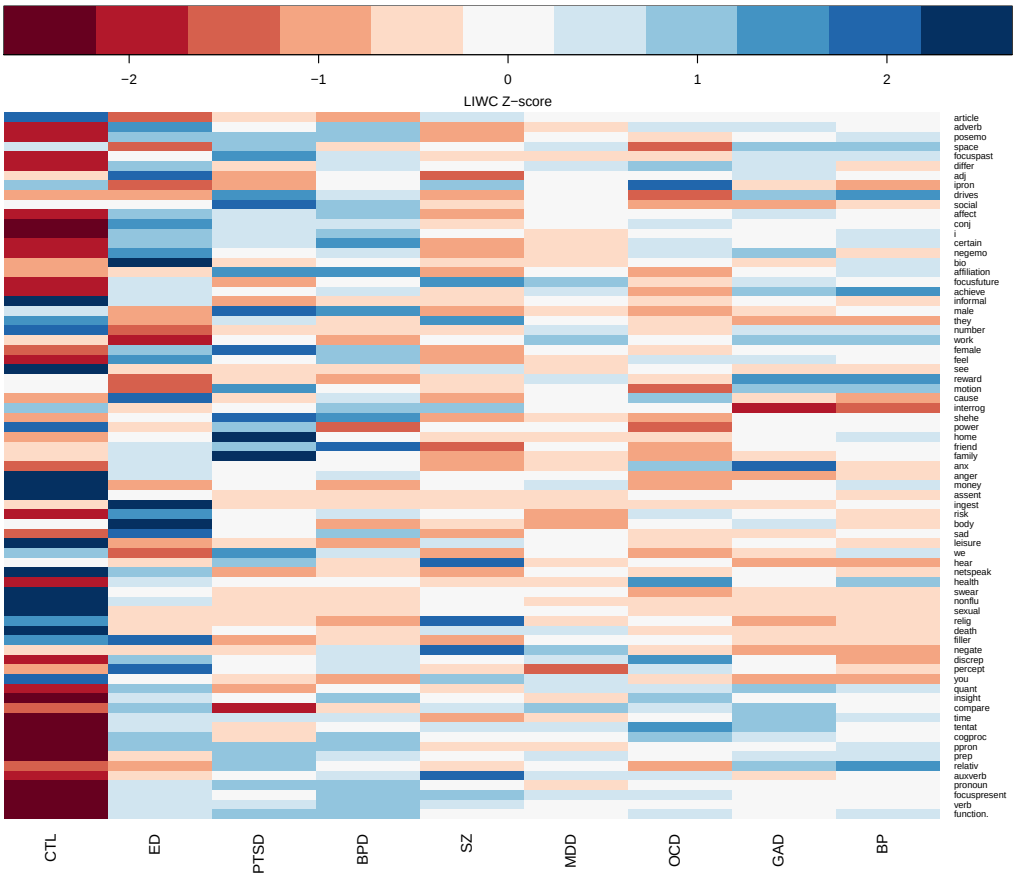


Figure 2: Scaled LIWC features across 8 diagnostic groups and control group.

diagnosis, based on an aggregated representation of their posts. In addition, we conduct post-level classification experiments with classic BERT fine-tuning settings. This experiment is done to discover the importance of global context in classification. Finally, in addition to these binary classifiers (diagnosis vs. control), we ensemble all our binary classification models as a multi-label classifier among different diagnostic groups, which is the ultimate goal for the application of this work.

For user-level classification, we select users not belonging to the co-morbidity group as a control group. To reduce the size of the data, we remove posts less than 50 characters long from both the mental group and the control group, as we hypothesize that these may not provide enough information for classification. Also we do not include control users with fewer than 20 non-sensitive posts. When pairing with mental health users, We select control users who have a similar total number of posts, who do not have mental health sensitive posts, and who have at least some subreddit overlap with the mental health users, as described by Cohan et al. (2018b). We consider two experimental settings: for the **CL** (clean) experiments we exclude all mental health sensitive posts for mental health users, and for the **UNCL** (unclean) experiments we include these posts for their corresponding users. The intuition is that under the **UNCL** setting, our model should be able to make predictions based on some explicit semantic triggers, thus resulting in better performance. However, under the **CL** setting, the model may rely on underlying syntactic differences that may generalize better than explicit semantic features.

Below we describe the Attention-Based model and the REALM model that we adapt for this work.

### 5.1 Attention-Based Model

A direct solution to the scalability issue with this data is to restrain the gradient update in a model to a small portion of the model parameters. We propose to use pre-trained BERT model from Hugging Face (Wolf et al., 2019) to encode every post;

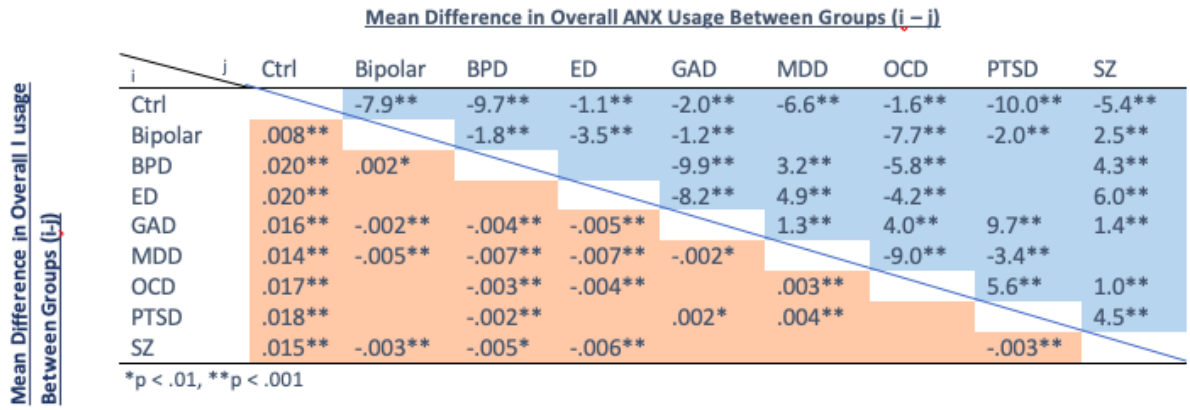


Figure 3: LIWC analysis across 8 diagnostic groups and control group for anxiety (ANX, in blue) and singular personal pronoun usage (I, in orange). Only significant results are displayed in this table.

we then averaged representation of all positions as a pooling result to build an attention-based classifier (Bahdanau et al., 2014; Sutskever et al., 2014) over all the post-level representations for a single user[3]. This resembles the settings of many "probing tasks" (Hewitt and Manning, 2019) used to investigate whether BERT embeddings encode useful linguistic information about a user's mental health condition.

### 5.2 REALM-like Models

Guu et al. (2020) propose Retrieval-Augmented Language Model pretraining to augment a pre-trained LM as a textual knowledge retriever. To tackle the scalability issue of retrieving over large corpora, a retrieval encoder parameterized by $\theta$ tuned over their top-$k$ retrieval results is used to encode all documents in the textual knowledge corpus. Guu et al. (2020) shows by gradient analysis that a document $z$ will receive a positive update if the estimated probability of a correct answer $y$ based on $z$ is better than the expectation over all documents in the textual knowledge corpus. To adapt this REALM model to our task, we reformulate our classification problem as a "retrieve-then-predict" pipeline similar to Open Domain Question Answering (ODQA). Specifically, given a user's total set of posts $\mathcal{Z}$, we first select the top-$k$ posts $\{z_1, \ldots, z_k\}$ that are most helpful in predicting the user's mental health condition and we base our prediction only on these posts. Unlike in ODQA we have a question $x$ that can be utilized for relevant document selection, so we now use a trainable attention head to calculate the retrieval probability $p(z)$. Thus the probability of a user having condition $y$ can be factorized as:

$$p(y) = \sum_{z \in \mathcal{Z}} p(y|z)p(z) \tag{1}$$

where

$$p(z) = \frac{h^T \text{Embed}_{\text{doc}}(z)}{\sum_{z'} h^T \text{Embed}_{\text{doc}}(z')} \tag{2}$$

Here, $\text{Embed}_{\text{doc}}(\cdot)$ is implemented as a BERT-style transformer parameterized by $\theta$ and $p(y|z)$ is implemented as a BERT-based classifier parameterized by $\phi$. When training, we first index all user posts with $\text{Embed}_{\text{doc}}(z)$ using our model $\theta$, and jointly tune $\theta$ and $\phi$ and $h$ on the top-$k$ user posts w.r.t. $p(z)$. For every several epochs, we re-index all posts with tuned parameter $\theta'$.

### 5.3 Experimental Settings

For REALM-like models we update the user corpus index every 5 epochs. At every step we use the top 10 documents to tune the model for each user, and we set the learning rate for the attention-based classifier to 1e-3, and the learning rate for BERT parameters to 1e-5. For the attention-based model, we set the learning rate for the classifier to 1e-3 and keep the BERT parameters frozen. For our post-level classification model we set the learning rate in the same way as in the REALM-like model. Note that when fine-tuning the BERT-based model, we pool the sentence representation with $[CLS]$ token, unlike the non-tunable model (BERT-ATT)

[3]Note that we are not using the $[CLS]$ (start sequence token) as the pooling result. This is because in the pretraining model it is used for next sentence prediction and thus is not an ideal semantic representation. The specific structure of BERT-ATT model does not allow the representation to be tuned.

where we average across all positions to get the pooling result. In all cases except for our LIWC-feature-based logistic regression model, we use a held-out development set for model selection; for LIWC-based regression we run a parameter grid search using cross-validation on the training set. For the multi-label classification experiments, we ensemble the best model set under the **CL** setting as the multi-label classifier.

## 6 Mental Health Detection Results

In this section we present the results for user-level and post-level binary classification, and for the multiclass ensemble classification.

### 6.1 User-Level Classification

Table 3 shows the user-level binary classification results, comparing the BERT-attention model, REALM model, and the LIWC logistic regression model, under both **CL** and **UNCL** settings. We find that, under the **UNCL** setting, the REALM-like model consistently achieves the highest accuracy for all diagnostic groups. Under the **CL** setting, tuning a classifier over the original BERT representation achieves better results for all groups. This is probably because the REALM model makes its predictions using only the top-10 segments retrieved from all of a user's posts, while the BERT-ATT model is able to attend to all the posts at once. This result aligns well with the intuition that linguistic traits for mental health conditions should be global, and may be more difficult to determine from a small portion of posts – especially when posts containing sensitive keywords are removed. The BERT-ATT model performs best for the bipolar category (CL-F1: .879; UNCL-F1: .931) for which we have the largest user group, indicating the importance of obtaining large scale datasets for the success of deep mental health detection. In all cases, our results suggest that contextualized representation is a better feature for mental health prediction compared with LIWC features, but is also more likely to model shallow semantic traits.

### 6.2 Post-Level Classification

To create a balanced set comparable to user level classification, we sample 50,000 mental group posts and 50,000 control group posts as the training set, and 5,000 + 5,000 posts for dev and test. Table 4 shows the post-level binary classification results. This post-level classification result ranges from an F1 of .596 for MDD to an F1 of .736 for ED, substantially lower than the user-level classification performance. This suggests that linguistic signals related to mental health problems do not appear in all posts of a mental group user. However, model performance exhibits similar trends when we compare post-level with user-level classifications LIWC features, indicating that the ED subset is the easiest and MDD the hardest: this result may mean that linguistic traits for ED have a broader coverage among user posts while for MDD the scope is probably smaller. This is consistent with the results reported by Coppersmith et al. (2015a).

### 6.3 Multi-label Model Ensemble

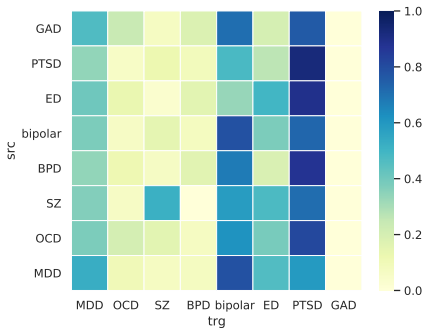


Figure 4: Multilabel ensemble experiments result. The overall result is $F-1_{micro} = 0.2175$ and $F-1_{macro} = 0.195$. Each cell representing the portion of users with gold label in src and predicted label in trg by our ensemble classifier.

As the BERT-ATT model performs best under **CL** setting, we ensemble all BERT-ATT model as the multi-label classifier. Again, to create a balanced testing set, we sample 100 users from each condition group's test set. We then predict the user condition by selecting the label with the highest score from the model. With this naive ensemble method we achieve $F - 1_{micro} = 0.2175$ and $F - 1_{macro} = 0.195$. The fact that these results are only slightly above the random baseline (.125) indicates that, though under binary settings deep contextualized word representation is a strong feature, the model is not well calibrated, (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005) as is often the case for modern deep networks (Guo et al., 2017). To see whether there are any identifiable patterns in the errors, we plot the prediction heatmap for the multi-label classification task, as shown in Figure 4. We find that there is a discrepancy of confidence between dif-

| | | CL | | | UNCL | | |
|---|---|---|---|---|---|---|---|
| | | **LIWC** | **BERT-ATT** | **REALM** | **LIWC** | **BERT-ATT** | **REALM** |
| **SZ** | Acc. | 0.668 | **0.808** | 0.727 | 0.757 | 0.885 | **0.97** |
| | F-1 | 0.685 | **0.812** | 0.766 | 0.775 | 0.898 | **0.973** |
| **BPD** | Acc. | 0.729 | **0.884** | 0.822 | 0.792 | 0.89 | **0.995** |
| | F-1 | 0.716 | **0.875** | 0.82 | 0.782 | 0.877 | **0.995** |
| **PTSD** | Acc. | 0.703 | **0.872** | 0.712 | 0.764 | 0.892 | **0.979** |
| | F-1 | 0.694 | **0.877** | 0.728 | 0.758 | 0.885 | **0.978** |
| **ED** | Acc. | 0.75 | **0.873** | 0.825 | 0.843 | 0.877 | **0.99** |
| | F-1 | 0.732 | **0.882** | 0.838 | 0.85 | 0.887 | **0.99** |
| **MDD** | Acc. | 0.67 | **0.833** | 0.822 | 0.702 | 0.91 | **0.965** |
| | F-1 | 0.663 | **0.843** | 0.819 | 0.719 | 0.908 | **0.965** |
| **GAD** | Acc. | 0.799 | **0.834** | 0.758 | 0.845 | 0.855 | **0.988** |
| | F-1 | 0.789 | **0.799** | 0.764 | 0.835 | 0.847 | **0.989** |
| **OCD** | Acc. | 0.721 | **0.865** | 0.746 | 0.815 | 0.884 | **0.973** |
| | F-1 | 0.71 | **0.872** | 0.707 | 0.807 | 0.875 | **0.974** |
| **Bipolar** | Acc. | 0.699 | **0.872** | 0.82 | 0.778 | 0.926 | **0.983** |
| | F-1 | 0.692 | **0.879** | 0.813 | 0.773 | 0.931 | **0.982** |

Table 3: User-level BERT classification results. The best result for a mental group under specific settings is in bold.

| | **Accu.** | **F-1** |
|---|---|---|
| SZ | 0.628 | 0.614 |
| BPD | 0.689 | 0.689 |
| PTSD | 0.630 | 0.577 |
| ED | 0.708 | 0.736 |
| MDD | 0.567 | 0.596 |
| GAD | 0.675 | 0.683 |
| OCD | 0.627 | 0.65 |
| Bipolar | 0.598 | 0.615 |

Table 4: Post-level BERT Classification Results

ferent models, and this confidence neither strongly correlates with training data size nor with binary classification performance. Though we observe that cells on the main diagonal in general have a darker shade indicating a promising separation of feature sets that are useful in identifying their designated condition, the mislabeling distribution has little resemblance to the comorbidity distribution characterized in Figure 1. Further experimentation is needed to improve the multiclass ensemble classification performance.

## 7 Conclusions and Future Work

In this paper we collect and analyze a large scale dataset of social media posts from users various mental health conditions. We analyze linguistic characteristics of the posts, directly comparing the features of the various conditions. We build strong classification models based on deep contextualized representations and demonstrate that they outperform the LIWC feature based logistic regression baseline by a large margin. Although the LIWC feature representation is not as useful for classification, it is a useful representation for analysis of posts to gain insight about the differences between groups. Our experimental results show that linguistic traits for mental health detection are more easily recognized at the user-level and thus effectively aggregating post-level signals is crucial to accurate prediction. Also, we find that these contextualized representations rely heavily on semantic content and always perform better when semantic indicators are obvious. We also show that the prediction scores of our classification models, even the accurate ones, are not well calibrated and thus are not an accurate uncertainty estimator of mental health risk. These results call for a more interpretable model for mental health detection. Future research may look into the direction of learning better deep features and exploring additional classification paradigms to further improve performance for this impactful problem.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378. ACM.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. # whywetweetmh: understanding why people use twitter to discuss mental health problems. *Journal of medical Internet research*, 19(4).

Victoria Bird, Preethi Premkumar, Tim Kendall, Craig Whittington, Jonathan Mitchell, and Elizabeth Kuipers. 2010. Early intervention services, cognitive–behavioural therapy and family intervention in early psychosis: systematic review. *The British Journal of Psychiatry*, 197(5):350–356.

Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 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.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018a. 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*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018b. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

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 *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in twitter. In *Eighth international AAAI conference on weblogs and social media*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Giovanni De Girolamo, J Dagani, R Purcell, A Cocchi, and PD McGorry. 2012. Age of onset of mental disorders and use of mental health services: needs, opportunities and obstacles. *Epidemiology and psychiatric sciences*, 21(1):47–57.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Elizabeth Highton-Williamson, Stefan Priebe, and Domenico Giacco. 2015. Online social networking in people with psychosis: a systematic review. *International Journal of Social Psychiatry*, 61(1):92–101.

Yen-Hao Huang, Lin-Hung Wei, and Yi-Shin Chen. 2017. Detection of the prodromal phase of bipolar disorder from psychological and phonological aspects in social media. *arXiv preprint arXiv:1712.09183*.

Andrew Hutchinson. 2018. Reddit now has as many users as twitter, and far higher engagement rates. https://www.socialmediatoday.com/news/reddit-now-has-as-many-users-as-twitter-and-far-higher-engagement-rates/521789/. Accessed: 2019-03-10.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 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*, pages 11–20.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. Quick and (maybe not so) easy detection of anorexia in social media posts. In *CLEF (Working Notes)*.

Andrea Murru and Bernardo Carpiniello. 2018. Duration of untreated illness as a key to early intervention in schizophrenia: a review. *Neuroscience letters*, 669:59–67.

Nona Naderi, Julien Gobeill, Douglas Teodoro, Emilie Pasche, and Patrick Ruch. 2019. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *Proceedings of the CLEF 2019 Workshop*.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

World Health Organization et al. 2014. *Global status report on noncommunicable diseases 2014*. WHO/NMH/NVI/15.1. World Health Organization.

Vikram Patel, Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, Pamela Y Collins, Janice L Cooper, Julian Eaton, et al. 2018. The lancet commission on global mental health and sustainable development. *The Lancet*, 392(10157):1553–1598.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

James W Pennebaker, Tracy J Mayne, and Martha E Francis. 1997. Linguistic predictors of adaptive bereavement. *Journal of personality and social psychology*, 72(4):863.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. *arXiv preprint arXiv:1811.04655*.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Janet Treasure and Gerald Russell. 2011. The case for early intervention in anorexia nervosa: theoretical exploration of maintaining factors. *The British Journal of Psychiatry*, 199(1):5–7.

Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136.

Daniel Vigo, Graham Thornicroft, and Rifat Atun. 2016. Estimating the true global burden of mental illness. *The Lancet Psychiatry*, 3(2):171–178.

Elizabeth Reisinger Walker, Robin E McGee, and Benjamin G Druss. 2015. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. *JAMA psychiatry*, 72(4):334–341.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

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

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83.