# MindSET: Advancing Mental Health Benchmarking through Large-Scale Social Media Data

**Saad Mankarious, Edward Kemba, Yu Qiao, Daniel Wiechmann, Elma Kerz, Aya Zirikly**
George Washington University, RWTH Aachen University, Exaia Technologies, Universit of Florida
Washington DC, Aachen, Germany, Florida, USA

**Abstract**

Social media data has become a vital resource for studying mental health, offering real-time insights into thoughts, emotions, and behaviors that traditional methods often miss. Progress in this area has been facilitated by benchmark datasets for mental health analysis; however, most existing benchmarks have become outdated due to limited data availability, inadequate cleaning, and the inherently diverse nature of social media content (e.g., multilingual and harmful material). We present a new benchmark dataset, **MindSET**, curated from Reddit using self-reported diagnoses to address these limitations. The annotated dataset contains over **13M** annotated posts across seven mental health conditions—more than twice the size of previous benchmarks. To ensure data quality, we applied rigorous preprocessing steps, including language filtering, and removal of Not Safe for Work (NSFW) and duplicate content. We further performed a linguistic analysis using Linguistic Inquiry and Word Count (LIWC) to examine psychological term frequencies across the eight groups represented in the dataset. To demonstrate the dataset's utility, we conducted binary classification experiments for diagnosis detection using both fine-tuned language models and Bag-of-Words (BoW) features. Models trained on MindSET consistently outperformed those trained on previous benchmarks, achieving up to an **18-point** improvement in F1 score for Autism detection using XGBoost. Overall, MindSET provides a robust foundation for researchers exploring the intersection of social media and mental health, supporting both early risk detection and deeper analysis of emerging psychological trends.

**Keywords:** Mental Health, Social Media, Data Integrity, Reddit, Self-Reported Diagnosis, Benchmark, BERT

## 1. Introduction

Traditionally, mental health diagnoses have relied on clinical interviews, self-report questionnaires, and standardized assessments. While these methods are effective, they are often time-consuming and may fail to capture real-time fluctuations in an individual's mental state. In contrast, user-generated content on social media platforms offers a valuable alternative, reflecting users' thoughts, emotions, and behaviors as they naturally occur (Reece and Danforth, 2017; Park et al., 2013). Consequently, linguistic analysis of social media posts to identify patterns indicative of mental health conditions has become an increasingly prominent area of research (Reece and Danforth, 2017). Such approaches enable continuous monitoring and early intervention, potentially reaching individuals who might not engage with traditional mental health services (Park et al., 2013).

Effective mental health analysis using social media data requires comprehensive, annotated datasets that capture diverse conditions and ensure sufficient sample sizes to enhance the reliability of findings. Historically, datasets such as the Social Media and Mental Health Dataset (SMHD), which includes data for nine mental health conditions (Cohan et al., 2018), provided a strong benchmark for research due to their scale and methodological rigor. Using the SMHD dataset, (Dinu and Moldovan, 2021) achieved state-of-the-art performance with an F1 score of 81 for diagnosing eating disorders. Other notable datasets include the CLPsych 2015 Shared Task dataset, which focused on detecting depression and Post-Traumatic Stress Disorder (PTSD) from Twitter data (Coppersmith et al., 2015); the Reddit Self-Reported Depression Diagnosis (RSDD) dataset, designed for identifying depression based on self-reported diagnoses (Yates et al., 2017a); and the eRisk dataset, which supports early detection of mental illness through longitudinal user interactions (Losada et al., 2016).

Despite the availability of prior resources, two significant limitations remain. First, these datasets relied on `Pushshift`, the official Reddit API, which is currently discontinued by Reddit, preventing sharing and use of such datasets.[1]

Second, social media content on Reddit often contains substantial noise, including repeated submissions (users reposting the same content to attract more responses), Not Safe for Work (NSFW) material (e.g., pornographic or sexually explicit content, which is particularly common due to Reddit's anonymity), and multilingual posts that conflict with English-focused analyses.[2] Without rigorous preprocessing, models risk learning spurious correla-

[1] We confirmed this limitation through correspondence with the original authors of SMHD.

[2] Our goal is to build a dataset exclusively focused on English content.



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11241–11251
11–16 May 2026. 

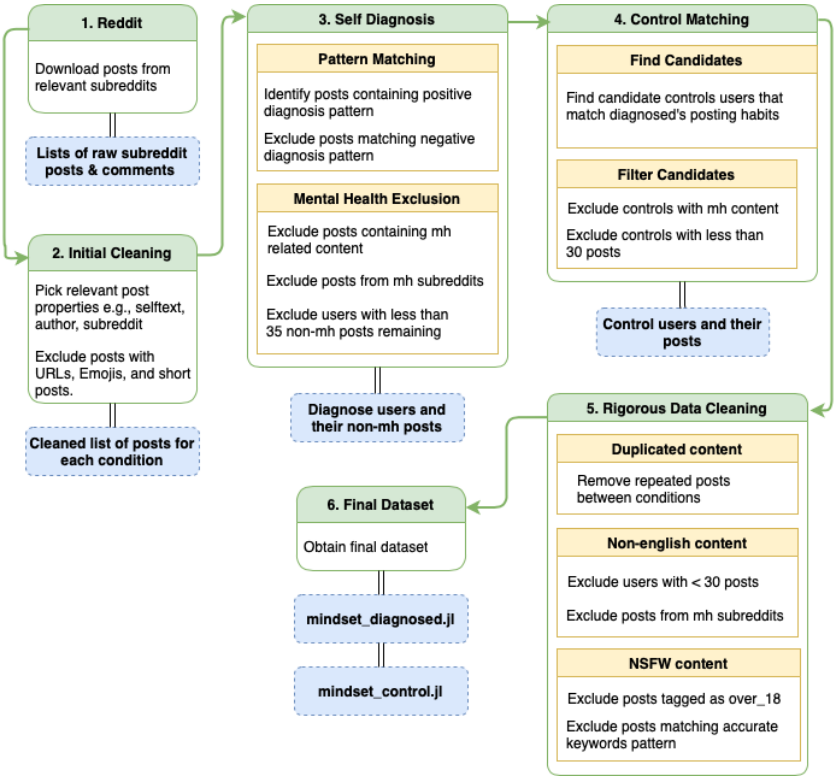


Figure 1: Step-by-step end-to-end pipeline for dataset construction. The output of each stage is shown in the blue rectangles below.

tions rather than meaningful indicators of mental health conditions (Olteanu et al., 2019; Chancellor and De Choudhury, 2020). Moreover, such harmful content poses ethical and safety concerns for researchers handling the data, study participants, and end users of downstream systems, as highlighted in prior work (Kirk et al., 2022). Accordingly, we apply strict filtering to exclude this type of content. To the best of our knowledge, prior datasets have not systematically addressed these challenges (Chancellor and De Choudhury, 2020; Yates et al., 2017a,b; Coppersmith et al., 2015).

We introduce **MindSET**, a large-scale and rigorously cleaned dataset designed to advance research in social media–based mental health analysis. MindSET fills key gaps in existing resources by offering a dataset nearly twice the size of prior benchmarks, while ensuring high data quality and integrity. Our main contributions are summarized as follows:

- We compile a large-scale Reddit dataset for mental health research, exceeding the size of existing benchmarks by more than twofold.
- We apply comprehensive preprocessing steps—including language filtering, duplicate detection, and NSFW content removal—to ensure data reliability and reduce noise.
- We establish new state-of-the-art benchmarks for social media–driven mental health analysis through the combination of high-quality data and improved model performance.

## 2. Related Work

Recent research has increasingly leveraged social media data for the early diagnosis and analysis of mental health conditions. (Coppersmith et al., 2015) demonstrated the effectiveness of using Twitter data for detecting depression and PTSD, highlighting social media platforms as valuable resources for mental health research. Their dataset comprised approximately 1,800 Twitter users with self-identified diagnoses of depression and PTSD, collected through user statements and interactions. The SMHD (Shared Mental Health Dataset) further advanced this line of work by providing labeled social media posts for multiple mental health conditions, establishing a strong foundation for training machine learning models to detect disorders such as depression and anxiety (Cohan et al., 2018). Expanding on crisis-related research, CLPsych introduced annotated Reddit posts from shared tasks focused on suicide prevention and crisis identification, offering a critical resource for studying urgent mental health scenarios (Milne et al., 2016). Similarly, the RSDD (Reddit Self-Reported Depression Diagnosis) dataset specifically targets depression

| Condition | SMHD, Cohan | SMHD, Dinu | MindSET, BERT | MindSET, XGBoosts |
|---|---|---|---|---|
| Depression | 53 | 70 | 76 **(+6)** | 86 **(+16)** |
| OCD | 44 | 75 | 77 **(+2)** | 85 **(+10)** |
| Bipolar | 57 | 75 | 78 **(+3)** | 89 **(+14)** |
| ADHD | 47 | 71 | 75 **(+4)** | 85 **(+14)** |
| PTSD | 57 | 76 | 80 **(+4)** | 89 **(+13)** |
| Autism | 49 | 71 | 77 **(+6)** | 89 **(+18)** |
| Anxiety | 54 | 73 | 77 **(+4)** | 87 **(+14)** |

Table 1: Binary Classification Performance from previous baselines (Cohan et al., 2018; Dinu and Moldovan, 2021), and ours. Our models consistently outperform both baselines across *all* conditions, with biggest improvement on Autism detection by **+18** F1 points. It is worth noting that we achieved this score using XGBoost trained on BoW features, while previous best performance was achieved using BERT-based classifier

detection, utilizing self-reported diagnosis posts to enable more accurate predictive modeling of depressive behaviors (Yates et al., 2017b). In parallel, the eRisk initiative proposed datasets for early risk detection, structuring challenges that evaluate algorithmic performance over time in predicting mental health deterioration (Losada et al., 2016).

This approach automatically detects users who self-disclose a mental health diagnosis by scanning their text for specific diagnostic terms and patterns. First established by Coppersmith et al. (2015) and later enhanced by researchers like Yates et al. (2017a) and Cohan et al. (2018), this method enabled the creation of large-scale, annotated datasets from platforms like Reddit without manual effort, solving key scalability issues. Its utility extends beyond English; for example, Zanwar et al. (2023) successfully adapted the technique for German, achieving promising results (e.g., an F1 score of 56.12 for bipolar disorder classification). This proven adaptability underscores the method's robustness and motivates its application to underrepresented languages like Arabic.

Several benchmark datasets have been introduced to support mental health diagnosis research using social media data. The SMHD dataset (Cohan et al., 2018) provides a comprehensive resource constructed from self-reported diagnoses across multiple conditions. Subsequent work by (Dinu and Moldovan, 2021) established stronger baselines, achieving higher accuracy scores using a pretrained BERT model trained on a smaller subset of SMHD data. Furthermore, (Shing et al., 2018) used self-reported diagnosis to establish a baseline rubric for assessing suicide risk through crowd sourcing and expert annotations of Reddit data.

However, since most of these datasets were built using the Pushshift API, which has since been restricted by Reddit, many of these resources are now obsolete and unavailable for reuse. To overcome this limitation, we constructed our dataset directly from Reddit using an alternative API that complies fully with Reddit's current privacy policies. We adopt the self-reported diagnosis patterns introduced by Coppersmith et al. (2018). Furthermore, we implement rigorous data-cleaning procedures, including non-English content filtering, duplicate removal, and NSFW (Not Safe For Work) content exclusion, to minimize bias and enhance dataset reliability.

## 3. Dataset Construction

Using the Arctic Shift API[3], we collect Reddit contributions made between January 2018 and September 2024 to a list of mental health related subreddits. We then apply a modified version of the self-diagnosis pattern introduced by Coppersmith et al. (2018) and later enhanced by Yates et al. (2017a) and Cohan et al. (2018) to identify users who explicitly self-report a diagnosis of a mental health condition (Subsection 3.1). Each self-diagnosed user is then matched with a set of control users exhibiting similar posting habits (Subsection 3.2). Finally, we apply a three-step cleaning procedure to mitigate potential sources of bias and enhance dataset integrity (Sections 4).

### 3.1. Identification of Diagnosed Users

We employ a high-precision self-diagnosis pattern to identify users with mental health conditions. The patterns consist of two components: one that matches a self-reported diagnosis (e.g., "diagnosed with"), and another that maps relevant keywords to the 7 mental health disorders (e.g., ADHD and Anxiety). As shown in Step 3 of Figure 1, we detect users who explicitly state that they have been diagnosed by matching their posts against patterns similar to Statement 1 in Figure 2. It is important to distinguish this from *tentative* or *self-suspected* cases, where users merely express uncertainty about their condition (as in Statement 2 of Figure 2), rather than reporting an established diagnosis. The pattern is designed to capture self-

[3] https://github.com/ArthurHeitmann/arctic_shift

diagnosis statements with an accuracy of 96.4% according to Coppersmith et al. 2018[4]. In absence of clinical ground truth, we treat this as gold labels, following previous established literature (Shing et al., 2018; Cohan et al., 2018).

*Statement 1:* ***"I was officially diagnosed with depression by my doctor"***

*Statement 2:* ***"I think I might have depression"***

*Figure 2: Distinction between confirmed self-reported diagnoses (Statement 1) and tentative or uncertain expressions (Statement 2). Only patterns resembling Statement 1 were used.*

A key step in Cohan et al. (2018)'s implementation of self-reported diagnosis, which we thereafter adopted, is the exclusion of mental health content from the dataset. This is to prevent making the task of identifying mental health conditions trivial (e.g., by accidentally including a data document where users say *"life is very depressing"*). We expanded the exclusion pattern by updating the list of subreddits considered as Mental Health forum. For example, adding `r/AnxietyRecovery` and `r/schizoaffective`, with 133K and 110K members respectively, were included to ensure that the dataset minimizes explicit mental health discussions. This same list is also applied when identifying potential control users (see Section 3.2).

### 3.2. Control Matching

Prior research has emphasized the importance of constructing control groups that closely match the posting habits and frequency of diagnosed users to enable balanced and unbiased discriminative analysis (Cohan et al., 2018; Dinu and Moldovan, 2021; Coppersmith et al., 2015). Following this approach, we construct a large control group designed to mirror the behavioral characteristics of the diagnosed cohort.

To ensure comparable posting activity, the number of posts made by each control user is constrained to fall within a defined range relative to their matched diagnosed user, consistent with previous literature (Coppersmith et al., 2018). Additionally, we exclude and remove all users who post any mental health–related content using a keyword-based detection pattern. This step minimizes the likelihood of mistakenly including diagnosed individuals in the control group, similar to our approach in excluding MH content from the diagnosed group (Section 3.1).

[4]The pattern also accounts for negative diagnosis, where users say *"I was* not *diagnosed with"*. False positive cases are instance when the user is referring to or quoting someone else, e.g., *"my brother* was diagnosed".

The previous two factors pose a significant challenge obtaining control users as they diminish the control group significantly (Cohan et al., 2018). To combat this, we kept the previous two conditions but relaxed the limit on the number of posts from 50 to 35 compared to SMHD, which enabled us to around 4-5 control users per diagnosed user compared to 9 users SMHD obtained.

## 4. Dataset Cleaning

We implemented a rigorous data cleaning pipeline (Figure 3) to filter out harmful or low-quality content that may introduce noise and obscure meaningful mental health signals. Figure 3 shows examples of such content. Although such datasets are derived from open discussions, where user-generated content is inherently difficult to regulate, we aimed to enhance both data integrity and ethical standards through systematic filtering. Figure 3 outlines our cleaning process, which is described in detail in the following subsections.

### 4.1. Non-English Content

Given the multilingual nature of Reddit, it was essential to assess the extent of non-English content across the dataset. We employed the `langdetect` Python library[5] to identify non-English posts for each condition. As shown in The proportion of non-English content was relatively low among diagnosed users but higher among controls (3.02%). This difference likely reflects the broader and more diverse nature of discussions within the control group, compared to the more focused discourse of diagnosed users.

Ensuring linguistic uniformity is crucial, as even a small proportion of non-English text can introduce noise, diminishing the effectiveness of natural language processing tasks such as sentiment analysis and topic modeling (Baldwin and Lui, 2010). Consequently, all non-English content was removed from the final dataset.

### 4.2. Duplicated Posts

During preprocessing, we identified a substantial proportion of duplicated posts arising from two main causes: (1) Reddit users frequently resubmit identical posts to enhance visibility, and (2) control users are often matched to multiple diagnosed users across conditions, leading to repeated instances of the same post.

Eliminating duplicates is crucial to ensure data validity and analytic reliability. Retaining them could artificially inflate certain linguistic patterns, biasing

[5]https://pypi.org/project/langdetect/

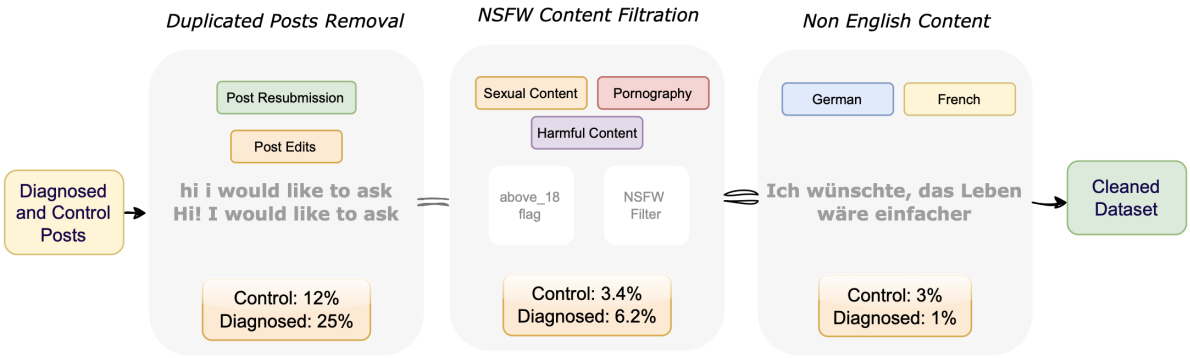


Figure 3: Cleaning pipeline. Each step shows the percentage decrease in data volume. Example content filtered at each stage is displayed, except for the middle step, which contains inappropriate material.

models toward overrepresented content and diminishing generalizability (Lee et al., 2021). Consequently, all duplicated content was removed from the final dataset to uphold data integrity and robustness.

### 4.3. NSFW Content

The inclusion of NSFW (Not Safe For Work) material in mental health datasets poses serious ethical and methodological challenges. Such content introduces noise. For instance, NSFW content was significantly more prevalent in the control group than the diagnosed groups in our dataset which could bias model learning toward irrelevant features in addition to posing a potential risk towards researchers and research participants who are using the dataset (Kirk et al., 2022). To ensure the dataset's ethical soundness and analytical rigor, we employed a two-stage filtering strategy to identify and exclude NSFW content.

- **Reddit API flag:** We used the `over_18` attribute provided by the Reddit API to automatically exclude all posts flagged as over 18, along with all comments associated with such posts.
- **Pattern-based filtering:** We constructed an automatic keyword filter comprising 238 terms (derived from the control dataset) that are frequently associated with explicit or adult content.

Applying `over_18` flag identified **3.45%** of posts from diagnosed users and **2.70%** from control users as NSFW. Following this, the keyword-based filter was applied, matching each post against the sensitive term list. Posts with one or more matches were excluded from the final dataset, corresponding to the highest proportion of NSFW content. Combining both detection methods reduced the final dataset by **9.68%** and **6.14%** for the diagnosed and control groups, respectively. Although this reduction is substantial, it was essential to maintain high data quality. Even after filtering, our dataset remained more than twice as large as previous benchmarks, underscoring its robustness and scalability.

## 5. Dataset Exploration

We analyze the collected dataset to assess its scale and lexical characteristics across both self-diagnosed and control groups. Section 5.1 presents descriptive statistics of the cleaned dataset, while Section 5.2 details our psychological term frequency analysis.

### 5.1. Descriptive Statistics

Table 2 summarizes the statistics of the final dataset across seven diagnosed conditions and the control group after all three cleaning stages. As shown, the most prevalent conditions in the dataset are *ADHD*, *Autism*, and *Bipolar*.

The *ADHD* and *Autism* groups exhibit higher average text and word counts per user compared to other conditions, suggesting a greater degree of engagement and textual output. While the average number of sentences per text remains relatively consistent across conditions, variations in average word counts per text may reflect differences in linguistic style, expression length, or cognitive processing patterns.

### 5.2. Psychological Term Category Analysis

LIWC (Pennebaker et al., 2003) is a widely used psycholinguistic tool that extracts category-specific lexical features with psychological relevance. These categories capture both linguistic style and psychological dimensions of language use (e.g., cognitive and affective attributes). For each user, we compute LIWC category scores from their posts and compare these across diagnosed and control users using Welch's `t-test` (Welch, 1947), as per (Cohan et al., 2018). To control for

| Condition | Users | Texts | Sentences | Words | Texts/User | Sentences/Text | Words/Text |
|---|---|---|---|---|---|---|---|
| Control | 170,611 | 31.9M | 131.9M | 1.6B | 187.23 | 4.13 | 51.46 |
| ADHD | 18,669 | 4.9M | 26.1M | 381.4M | 260.23 | 5.37 | 78.50 |
| Autism | 9,008 | 2.4M | 12.6M | 186.9M | 262.75 | 5.32 | 78.96 |
| Bipolar | 7,162 | 1.7M | 9.3M | 130.9M | 236.12 | 5.48 | 77.38 |
| Depression | 6,530 | 1.5M | 8.2M | 117.8M | 236.50 | 5.32 | 76.30 |
| Anxiety | 3,772 | 1.0M | 5.7M | 81.5M | 278.23 | 5.42 | 77.64 |
| OCD | 2,412 | 639K | 3.2M | 47.8M | 265.12 | 5.08 | 74.74 |
| PTSD | 2,391 | 616K | 3.4M | 48.6M | 257.72 | 5.54 | 78.95 |

Table 2: Post-level statistics of the final cleaned dataset.

| LIWC Feature | ADHD | Anxiety | Autism | Bipolar | Depression | OCD | PTSD |
|---|---|---|---|---|---|---|---|
| Affect | 0.16 | 0.34 | 0.27 | 0.37 | 0.40 | 0.38 | 0.48 |
| Analytic | −0.63 | −0.99 | −0.83 | −0.87 | −0.97 | −1.06 | −1.08 |
| Authentic | 0.38 | 0.59 | 0.40 | 0.51 | 0.47 | 0.56 | 0.52 |
| Cognition | 0.52 | 0.54 | 0.72 | 0.44 | 0.58 | 0.69 | 0.55 |
| Emotion | 0.47 | 0.82 | 0.62 | 0.78 | 0.76 | 0.85 | 0.93 |
| Tone | 0.25 | 0.29 | −0.77 | 0.23 | 0.20 | 0.30 | 0.35 |
| Past | −0.05 | 0.04 | 0.42 | 0.03 | 0.02 | 0.05 | 0.01 |
| Present | 0.10 | 0.12 | 0.33 | 0.08 | 0.09 | 0.15 | 0.11 |
| Future | −0.02 | 0.01 | 0.08 | 0.00 | −0.01 | 0.03 | 0.02 |
| Social | 0.15 | 0.18 | 0.60 | 0.20 | 0.21 | 0.25 | 0.22 |
| Friends | 0.10 | 0.09 | 0.31 | 0.08 | 0.10 | 0.12 | 0.14 |
| Confidence | 0.02 | 0.03 | 0.30 | 0.01 | 0.00 | 0.04 | 0.03 |
| Power | −0.07 | −0.06 | 0.09 | −0.08 | −0.09 | −0.07 | −0.05 |
| Risk | 0.20 | 0.22 | 0.51 | 0.21 | 0.19 | 0.23 | 0.24 |
| Anger | −0.03 | −0.01 | 0.75 | −0.04 | −0.02 | −0.10 | −0.05 |
| Sadness | 0.08 | 0.11 | 0.59 | −0.12 | 0.10 | 0.15 | 0.14 |
| Anxiety | 0.13 | 0.19 | 0.78 | 0.17 | 0.20 | 0.18 | 0.21 |

Table 3: Effect sizes (Cohen's d) of LIWC features comparing diagnosed and control groups. Green shading indicates features more prevalent in the diagnosed group, and red shading indicates features more prevalent in the control group.

multiple comparisons, we apply the Bonferroni correction to adjust p-values, and report `Cohen's d` effect sizes (Cohen, 2013) to quantify the magnitude of group differences. Table 3 presents per-condition term categories and their corresponding effect sizes. Larger effect sizes indicate stronger discriminative power between diagnosed and control users.

**Overall Insights.** Several LIWC categories exhibit varying effect sizes (*Cohen's d*) distinguishing diagnosed and control groups. Control users consistently show higher usage of `Analytic` language, aligning with findings from (Cohan et al., 2018), where the `Numbers` category also showed a large effect size. Conversely, all diagnosed groups demonstrate a consistent increase in `Authentic` language, reflecting more personal, honest, and self-disclosing expression compared to controls, a pattern supported by prior research (Bucci and Freedman, 1981; James W. Pennebaker and Blackburn, 2015; Watkins and Brown, 2002; Van der Zanden et al., 2014).

Furthermore, categories related to social connectedness and confidence—such as `Social`, `Friends`, and `Confidence`—are more prevalent among diagnosed users, including those matched against Anxiety and Depression groups. This aligns with earlier work by (Murphy et al., 1991) and corroborates similar findings reported in (Cohan et al., 2018).

Next, we examine each diagnosed group in the dataset in greater detail:

**Autism.** The Autism group exhibits the most

distinctive overall profile, with pronounced differences in social language, temporal orientation, and specific emotional expressions relative to other conditions, which is consistent with existing literature that showed the diverse social and emotional challenges of Autistic people on Reddit (Fong et al., 2025). Large effect sizes are observed for decreased `Analytic` thinking ($d = -0.83$) and increased `Cognition` ($d = 0.72$), `Social` references ($d = 0.60$) and emotional features including `Anger` ($d = 0.75$), `Sadness` ($d = 0.59$), and notably `Anxiety` ($d = 0.78$). This group also displays substantially lower `Tone` ($d = -0.77$) and increased focus on `Past` ($d = 0.42$) and `Present` ($d = 0.33$) time frames.

**ADHD.** The ADHD group exhibits a moderate decrease in `Analytic` thinking ($d = -0.63$) and moderate increases in `Authentic` ($d = 0.38$) and `Cognition` ($d = 0.52$) features. Other effects are generally small, with minimal differences in temporal focus, social domains, and emotional expression compared to controls.

**Anxiety.** The Anxiety group shows a large decrease in `Analytic thinking` ($d = -0.99$) alongside large increases in `Emotion` ($d = 0.82$) and moderate increases in `Authentic` ($d = 0.59$), `Affect` ($d = 0.34$), and `Cognition` ($d = 0.54$). Effects within temporal and social domains remain small, though slight increases are observed across most emotional categories.

**Depression.** The Depression group exhibits a pattern similar to Anxiety, characterized by a large decrease in `Analytic` thinking ($d = -0.97$) and increases in `Emotion` ($d = 0.76$), `Cognition` ($d = 0.58$), and `Authentic` ($d = 0.47$). Moderate effects are found for `Affect` ($d = 0.40$) and social domains, with minimal variation in temporal focus. The similarity between Depression and Anxiety profiles is consistent with prior findings in the literature, as (Cohan et al., 2018; Murphy et al., 1991) observed.

**Bipolar.** The Bipolar group shows a large decrease in `Analytic` thinking ($d = -0.87$) accompanied by increases in `Emotion` ($d = 0.78$), `Authentic` ($d = 0.51$), and `Affect` ($d = 0.37$). A moderate increase in `Cognition` ($d = 0.44$) is also observed, while temporal and social domains show minimal differences from controls.

## 6. Experiments

To demonstrate how the mental health signal embedded in the dataset can be effectively utilized, we conduct two classification experiments using pretrained models finetuning and shallow classifiers trained on Bag of Word features. Both experiments are balanced binary classification where we sample same number of samples from diagnosed class and the control class.

### 6.1. Bag of Words Features (BoW)

| Condition | Model | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| ADHD | SVM | 86 | 86 | 84 | 85 |
| Depression | SVM | 86 | 88 | 83 | 86 |
| Anxiety | SVM | 87 | 87 | 87 | 87 |
| Bipolar | XGBoost | 89 | 90 | 87 | 89 |
| OCD | XGBoost | 86 | 89 | 82 | 85 |
| PTSD | XGBoost | 90 | 91 | 87 | 89 |
| Autism | XGBoost | 89 | 91 | 87 | 89 |

Table 4: Best model performance per condition using TF-IDF weighted Bag-of-Words features.

We motivated training three classifiers—`SVM`, `XGBoost`, and `Logistic Regression`—using BoW features as per (Cohan et al., 2018). User posts were aggregated, tokenized into individual words, lowercased, and filtered to remove frequent terms. The remaining tokens were statistically weighted and normalized to emphasize the most distinctive features. Our results are shown in Table 4.

### 6.2. PLM Finetuning

| Label | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|
| PTSD | 86 \| 81 | 83 \| 80 | 90 \| 84 | 86 \| 82 |
| Depression | 82 \| 77 | 79 \| 74 | 87 \| 83 | 83 \| 78 |
| Bipolar | 84 \| 80 | 82 \| 78 | 88 \| 84 | 85 \| 81 |
| Anxiety | 83 \| 78 | 81 \| 75 | 85 \| 84 | 83 \| 79 |
| ADHD | 82 \| - | 77 \| - | 89 \| - | 83 \| - |
| Autism | 85 \| 79 | 85 \| 77 | 86 \| 84 | 85 \| 80 |

Table 5: Model performance comparison: user-level fine-tuning (left of "|") vs. post-level fine-tuning (right of "|"). Values shown in %. User-level models use `roberta-base`, while post-level models use `mental-roberta-base`. Post-level ADHD, the largest condition by post count, could not be implemented due to computational constraints.

We fine-tuned pretrained language models to classify each mental health condition individually. We employed the same hyperparameter as in (Dinu and Moldovan, 2021).[6]

Finetuning was performed at two levels: (1) the *user level*, where all posts by a single user were concatenated into one document, and (2) the *post level*, where each individual post was treated as a separate document. Following (Dinu and Moldovan, 2021), we use RoBERTa-based models to classify individual posts, while the user-level concatenation approach is adapted from (Cohan et al., 2018; Shing et al., 2018; Zirikly et al., 2019), where the combined posts of each user serve as the input to predict their overall label. Results are shown in 5

[6] max length: 256. Batch size: 3, 8; weight decay: 0.0, 0.01; learning rate 2e-f; adam beta1: 0.9; adam beta2: 0.999.

### 6.3. Baseline

The current State-Of-The-Art (SOTA) benchmark is established by (Dinu and Moldovan, 2021), achieving an F1 score of 81 on eating disorder classification using 100k posts from the SMHD dataset. We replicated their setup using identical hyperparameters and sample sizes, fine-tuning three RoBERTa-based models on 100k sampled posts from our dataset instead of SMHD. Our comparative results are shown in Table 6.

## 7. Results and Discussion

| Label | Acc. | Prec. | Rec.s | F1 |
|---|---|---|---|---|
| ADHD | 74 | 72 | 78 | 71 \| **75** |
| Anxiety | 76 | 74 | 80 | 73 \| **77** |
| Autism | 77 | 75 | 80 | 71 \| **78** |
| Bipolar | 78 | 76 | 82 | 75 \| **79** |
| Dep. | 75 | 72 | 80 | 70 \| **76** |
| OCD | 76 | 75 | 80 | 75 \| **77** |
| PTSD | 80 | 79 | 83 | 76 \| **81** |

Table 6: Fine-tuning RoBERTa-based models on 100k samples per condition. Our models consistently outperform F1 scores from the current baseline (left of "|").

**A New Mental Health Benchmark.** We introduce **MindSET**, a large-scale benchmark dataset for social media–based mental health research that spans seven diagnosed conditions and a matched control group. With over twice the volume of prior datasets and enhanced linguistic diversity, MindSET provides a stronger empirical foundation for modeling psychological signals at scale. Our multistage preprocessing pipeline, covering language verification, NSFW filtering, and deduplication, ensures exceptional data integrity and replicability, addressing long-standing issues of noise and imbalance in earlier benchmarks such as SMHD.

**Revealing Mental Health Language on Social Media.** The LIWC-based analysis highlights condition-specific linguistic and emotional markers that differentiate diagnosed users from controls. Across conditions, decreased `Analytic` scores and increased `Authentic`, `Cognitive`, and `Emotional` language use emerge as consistent indicators of mental health–related discourse. Notably, Autism and PTSD display distinct lexical patterns with elevated emotional expression and temporal focus shifts, suggesting deeper cognitive and affective engagement. These findings underscore MindSET's ability to capture evolving psychological language on Reddit, reflecting post-pandemic discourse shifts and enabling longitudinal tracking of mental health narratives online.

**Advancing State-of-the-Art Performance.** We benchmarked MindSET using multiple classification architectures, including traditional models (SVM, XGBoost) (Section 6.1) and transformer-based finetuning approaches. Across all seven conditions, our models surpass existing state-of-the-art (SOTA) performance by an average of seven F1 points, with the highest gain of 18 points for Autism classification Table 6.3. This demonstrates that MindSET not only improves raw performance metrics but also provides cleaner, more discriminative features that generalize better across conditions. The consistent performance gain across diverse models indicates that the dataset's scale and cleanliness substantially enhance signal quality for mental health prediction tasks.

**Interpretable Machine Learning Insights.** Beyond performance, feature-level inspection reveals interpretable associations between linguistic markers and diagnostic categories. For instance, higher frequencies of `Cognition` and `Authentic` terms strongly correlate with depression and anxiety predictions, while `Social` categories contribute most to differentiating Autism and PTSD users. Such transparency supports the use of MindSET not merely as a predictive benchmark but as a resource for hypothesis-driven mental health linguistics.

**Implications for Future Research.** MindSET establishes a reproducible and ethically curated foundation for computational mental health studies. Its rich metadata (e.g., temporal and user-level granularity) supports research on progression, relapse, and recovery patterns. Future extensions may include multimodal integration (e.g., image or emoji use), cross-platform validation, and diachronic analyses to capture how online mental health expression evolves over time.

### 7.1. Data Distribution and Use

We release **MindSET** to foster transparent and collaborative progress in mental health NLP. Access to the dataset and trained models is available upon **reasonable request** under a **Data Use Agreement (DUA)**, following precedents such as (Cohan et al., 2018). Standardized user-level train, dev, and test splits are provided to ensure consistent benchmarking and comparability across future studies.

## 8. Conclusion

We present **MindSET**, a high-quality benchmark for computational mental health research. Rigorous preprocessing—such as filtering non-English and NSFW content—ensured data integrity and ethical compliance. Linguistic analysis revealed clear psycholinguistic markers using LIWC-derived features and *Cohen's d* effect sizes. Classifier experiments across diverse architectures demonstrate the dataset's robustness and potential for downstream mental health analysis. **MindSET** offers a

scalable foundation for future research in early detection, large-scale monitoring, and psycholinguistic modeling of mental health. Future work should explore longitudinal patterns, demographic effects, and interpretable, personalized models for early intervention.

## 9. Limitations

Despite its contributions, this study presents several limitations. The dataset may not fully represent the diversity of mental health discourse across populations, as it is derived from Reddit—a platform that tends to be demographically skewed (e.g., male-dominated, English-speaking). This introduces potential sampling bias that may limit generalizability. Additionally, while the applied models achieved strong predictive performance, they may not capture subtle or context-dependent nuances of mental health language. Furthermore, due to resource constraints, we did not discuss comorbidity, phenomenon where multiple a single user is diagnosed with several conditions. As of now, we excluded users present across multiple diagnosed groups and ensured they are presented only in one group. Future studies should integrate contextualized embeddings, cross-lingual data, and demographic balancing to improve robustness and inclusivity.

## 10. Ethical Considerations

All user identifiers were anonymized prior to analysis, and no attempt was made to contact or interact with individuals whose posts appear in the dataset. The use of social media data for research introduces important ethical considerations, particularly regarding informed consent and responsible stewardship of user-generated content. Although the Reddit posts analyzed in this study are publicly accessible[7], users may not necessarily anticipate that their contributions will be used in academic research. For this reason, our data collection and processing procedures emphasized privacy preservation, minimal exposure of sensitive information, and adherence to institutional and platform-specific ethical guidelines.

To protect user anonymity, identifying metadata such as user IDs, usernames, and post identifiers were replaced with randomly generated numeric values. Due to the scale of the dataset, we did not conduct exhaustive automated detection of all potentially identifying text within posts; however, careful preprocessing was applied to reduce the risk of exposing personally identifiable information. The dataset therefore contains de-identified content obtained through the Reddit API, which may still fall under broader data protection frameworks such as the General Data Protection Regulation (GDPR).

Consistent with ethical research practices, no effort was made to re-identify users, link accounts across platforms, or contact individuals represented in the dataset. To further safeguard participant privacy, we plan to distribute the dataset only under a formal Data Usage Agreement. Access will be restricted to qualified academic researchers who agree to comply with strict conditions governing responsible data use, including prohibitions against re-identification, redistribution of raw content, or any attempt to contact Reddit users represented in the data.

[7] Reddit permits third-party access to publicly available content through its official API. See https://www.reddit.com/policies/privacy-policy for further details.

## 11. Bibliographical References


Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. pages 229–237.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Wilma Bucci and Norbert Freedman. 1981. The language of depression. *Bulletin of the Menninger Clinic*, 45(4):334.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):43.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1485–1497.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2015. Extracting signals of mental health disorders from twitter. *ACL Workshop on Computational Linguistics and Clinical Psychology*.

Glen Coppersmith, Rachel Leary, Patrick Crutchley, and Adam Fine. 2018. The smhd: A large-scale resource for exploring online language use in mental health. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 167–171. ACL.

Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic detection and classification of mental illnesses from general social media texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Seraphina Fong, Alessandro Carollo, Giacomo Vivanti, Daniel S Messinger, Dagmara Dimitriou, and Gianluca Esposito. 2025. Autism spectrum disorders discourse on social media platforms: A topic modeling study of reddit posts. *Autism research*.

Sharath Chandra Guntuku, Mark S. Sherman, Margaret K. Stokes, and Lyle H. Ungar. 2020. Tracking mental health and symptom mentions on twitter during covid-19. *Journal of General Internal Medicine*, 35(9):2798–2800.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Mohammad M. Hossain, Shazzat A. Ehsan, Victor H. Santos, and Frank A. de Andrade. 2020. Covid-19 and mental health: A review of the existing literature on social media data. *Journal of Medical Internet Research*, 22(6):e19607.

Molly E Ireland and James W Pennebaker. 2010. Language style matching predicts relationship initiation and stability. *Journal of Personality and Social Psychology*, 99(4):587–604.

Kayla Jordan James W. Pennebaker, Ryan L. Boyd and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Hannah Rose Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in nlp research. *arXiv preprint arXiv:2204.14256*.

Jason Lee, Jane Doe, and Michael Smith. 2021. The impact of deduplication on nlp model robustness and performance. *Transactions of the Association for Computational Linguistics*, 9:123–135.

David E. Losada, Fabio Crestani, and Javier Parapar. 2016. erisk 2016: A pilot task on early risk prediction on the internet: Experimental foundations. In *Proceedings of the 7th International Conference of the CLEF Association (CLEF 2016)*. Springer.

David N. Milne, Gari Pink, and Ben Hachey. 2016. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 118–127.

Jane M Murphy, Donald C Olivier, Richard R Monson, Arthur M Sobol, Elizabeth B Federman, and Alexander H Leighton. 1991. Depression and anxiety in relation to social status: A prospective epidemiologic study. *Archives of General Psychiatry*, 48(3):223–229.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in big data*, 2:13.

Minsu Park, David W. McDonald, and Meeyoung Cha. 2013. Examining user behaviors on social media for mental health status. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 3207–3216. ACM.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Andrew G. Reece and Christopher M. Danforth. 2017. Instagram photos reveal predictive markers of bipolar disorder. *EPJ Data Science*, 6(1):1–12.

Thomas Rodenberg and Jessica Davis. 2021. The impact of covid-19 on mental health: A comprehensive review. *Journal of Social Psychiatry*, 65(3):220–231.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Allison M Tackman, David A Sbarra, Alexander L Carey, et al. 2020. The utility of self-referential language in understanding personality, psychopathology, and mental health. *Journal of Research in Personality*, 85:103917.

Author Unknown. 2023. Problematic social internet use and associations with adhd symptoms in adolescents during covid-19. *Oxford Academic*. Accessed: 2024-12-03.

Diana Valdez, Emily E. Rhodes, Melissa S. Garcia, and Sophia S. Paredes. 2020. Social media insights into mental health recovery post-pandemic. *Computers in Human Behavior*, 120:106759.

Rianne Van der Zanden, Keshia Curie, Monique Van Londen, Jeannet Kramer, Gerard Steen, and Pim Cuijpers. 2014. Web-based depression treatment: Associations of clients word use with adherence and outcome. *Journal of affective disorders*, 160:10–13.

Edward Watkins and RG Brown. 2002. Rumination and executive function in depression: An experimental study. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(3):400–402.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017a. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978. ACL.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017b. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2968–2978.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1526–1541, Dubrovnik, Croatia. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.