MINDSET: A Novel Dataset for Mental Health Analysis

Anonymous ACL submission

Abstract

Social media data has become increasingly important in the study of mental health conditions, offering insights into real-time behaviors, thoughts, and emotions that traditional methods may not capture. Platforms such as Reddit underwent significant changes in usage patterns during and following the COVID-19 pandemic, impacting the ways people engage and communicate online. To capture and analyze these shifts, we developed a dataset from Reddit that focuses on social media data collected during this period. The dataset encompasses posts from seven mental health conditions, as well as a control group, offering a diverse view of users' language use over time. We employed an adapted version of a validated self-diagnosis pattern to accurately identify users who self-reported diagnoses. Through careful preprocessing, including language detection and duplicate content removal, we aimed to ensure a high level of data integrity. Machine learning models were subsequently trained on this refined dataset, demonstrating consistent performance across different conditions. This dataset serves as a foundation for researchers interested in exploring the evolving role of social media in mental health discourse, with the potential to support early detection and deeper analysis of mental health trends.

1 Introduction

Traditionally, mental health diagnoses relied on clinical interviews, self-report questionnaires, and standardized tests. While these methods are effective, they are time-consuming and may not capture real-time fluctuations in an individual's mental state. User-generated content on social media platforms offers an alternative, providing a valuable source of data that reflects users' thoughts, emotions, and behaviors in real-time (Reece and Danforth, 2017; Park et al., 2013). Linguistic analysis of social media posts to identify patterns indicative of mental health conditions has, therefore, become an increasingly prominent field of study (Reece and Danforth, 2017). Such tools can facilitate continuous monitoring and early intervention, potentially reaching individuals who may not engage with traditional mental health services (Park et al., 2013).

Effective analysis of mental health via social media requires comprehensive, labeled datasets that capture a diversity of 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 (Coppersmith et al., 2018), and the CLPsych 2015 Shared Task dataset, focused on depression and PTSD detection using Twitter data (Coppersmith et al., 2015), have been utilized. Additionally, the Reddit Self-Reported Depression Diagnosis (RSDD) dataset has been employed for depression detection based on self-reported diagnoses (Yates et al., 2017). The eRisk dataset also facilitates early detection of mental illness through structured user interactions over time (Losada et al., 2016).

We introduce the MINDSET dataset, which extends these efforts by addressing the need for a large-scale, labeled dataset comprising post-COVID data. By incorporating data from both during and after the pandemic and implementing stringent data-cleaning processes, MINDSET offers a reliable and current resource for understanding the mental health impacts of this period. Our work includes the following contributions:

- Collecting data created from 2019 onwards to capture post-pandemic mental health trends and behaviors.
- Updating the self-diagnosis matching pattern used in previous datasets like SMHD to en-

hance robustness of diagnosed users' data.

- Implementing rigorous data cleaning procedures, including non-English filtration and duplicate removal, to eliminate noise and boost dataset quality compared to previous datasets.

2 Related Work

Research has increasingly leveraged social media data for the early diagnosis of mental health conditions. Coppersmith et al. (2015) demonstrated the effectiveness of using Twitter data for detecting depression and PTSD, highlighting the potential of social media platforms as valuable resources for mental health research (Coppersmith et al., 2015). Their dataset included approximately 1,800 Twitter users with self-identified diagnoses of depression and PTSD, collected based on user statements and interactions. However, one limitation of using Twitter data is the character constraint; at the time, tweets were limited to 140 characters, which restricts the depth and context of the shared content.In contrast, platforms like Reddit allow for longer posts and detailed discussions, providing richer context and more comprehensive insights into users' mental health conditions and experiences. This enables a deeper linguistic and psycholinguistic analysis, making it a more suitable platform for building robust mental health datasets (Coppersmith et al., 2018).

The Social Media and Mental Health Dataset (SMHD) is another significant effort in utilizing social media for mental health research, focusing on Reddit as a data source. The dataset covers nine mental health conditions, such as depression, anxiety, PTSD, and bipolar disorder, providing a broad spectrum for analysis (Coppersmith et al., 2018). SMHD introduced a systematic method for identifying both diagnosed and control users, offering a substantial resource for the study of various conditions. However, while the dataset is large and comprehensive, it did not implement filtering mechanisms for duplicate or non-English content, which could introduce bias and affect the reliability of the results. This lack of data cleaning may lead to overrepresentation of certain posts and skew analysis, highlighting the need for more refined preprocessing techniques in future datasets.

Similar to SMHD (Coppersmith et al., 2018), we construct our dataset from Reddit using a self-reported diagnosis pattern. However, we extend this approach by implementing several enhancements to improve data quality and relevance. We update the self-diagnosis matching patterns to capture emerging terminology and trends within mental health communities, ensuring that our dataset reflects current discourse. Additionally, we introduce rigorous data-cleaning procedures, including filtering out non-English posts and removing duplicate content, to minimize bias and enhance the reliability of our findings.

3 Data

We are collecting raw data from Reddit using the Arctic Shift API [1]. After gathering posts from relevant subreddits we apply the self-diagnosis pattern (see Subsection 3.1) presented by (Coppersmith et al., 2018) to identify users who self-report as diagnosed with these conditions. The main subreddits used and the number of self-diagnosed users per condition are summarized in Table 1.

Table 1 reveals two additional key observations. First, there is significant variability in the number of raw posts required to identify a single diagnosed user across conditions. On average, 363 raw posts are needed to find one diagnosed user. The most efficient ratio is observed in *OCD*, with 127.66 posts per user, while the least efficient is seen in *PTSD*, with 820.51 posts per user. This suggests that it is more challenging to identify diagnosed users for conditions like *PTSD*, likely because users may not explicitly state their diagnosis or may discuss their symptoms in broader, less specific terms. Understanding these patterns is crucial for refining data collection methods to improve the accuracy and efficiency of user identification across various conditions.

Additionally, the reduction in the number of diagnosed users before and after exclusion reflects our strategy of setting a minimum of 30 posts per qualified user. This compromise reduces the total number of users but ensures a higher post count per user, improving the quality of the dataset.

3.1 Self-Diagnosis

We employ an enhanced version of the high-precision self-diagnosis pattern introduced by Coppersmith et al. (2018). This method identifies users who explicitly self-report having a mental disorder by matching posts against key phrases such as *"I*

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

Table 1: User and post counts at different stages of the data collection process. The total number of raw posts inspected is shown, with the final post count in parentheses (this may be higher than the raw post count as it includes additional user submissions fetched at a later stage). Also listed are the number of diagnosed users before and after applying exclusion criteria, the average number of posts per qualified user, and the total number of tokens (words) for each condition.

Condition	Subreddits	Posts (Final)	Users (After)	Posts/User	Tokens
Bipolar	*r/bipolar* *r/bipolar2*	2.1M (1.8M)	9.8K (7.2K)	291	144.41M
Depression	*r/depression* *r/SuicideWatch*	4.6M (1.6M)	13.0K (6.7K)	209	129.78M
Anxiety	*r/Anxiety* *r/SocialAnxiety*	4.7M (1.1M)	6.2K (3.9K)	233	89.56M
OCD	*r/OCD* *r/OCDsupport*	0.6M (0.7M)	4.7K (2.4K)	198	52.48M
PTSD	*r/PTSD* *r/trauma*	3.2M (0.7M)	3.9K (2.4K)	206	53.87M
ADHD	*r/ADHD* *r/ADHDWomen*	4.6M (5.2M)	29K (19.5K)	260	419.23M
Autism	*r/autism* *r/aspergers*	8.9M (2.5M)	12.4K (9.4K)	253	205.96M

was diagnosed with." The pattern focuses on capturing concise self-diagnosis statements, typically within a 40-character limit, to accurately identify direct admissions of mental health conditions. This self-diagnosis pattern identifies diagnosed users with an accuracy of 96.4 percent (Coppersmith et al., 2018). As per Coppersmith et al., we make sure any mental health related content of the diagnosed users is filtered out because otherwise the classification task will be superficially easy.

We enhanced Coppersmith et al.'s self-diagnosis pattern by updating the list of subreddits used to exclude mental health content for diagnosed users. We included 32 subreddits over the exiting 78 to consider when deciding to exclude mental health content (same list is also used for excluding candidate control users, see section 3.2). The new subreddits we added, such as *r/AnxietyRecovery* and *r/schizoaffective*, with 133K and 110K users respectively, help ensure that the dataset does not contain explicit mental health discussions.

Coppersmith et al. excludes users who have less than 50 Reddit contributions to make sure enough content remains after removing their mental health related content. To ensure only users with substantial social media activity are incl uded, we added another threshold of 30 posts for diagnosed users to have remaining after exclusion. This additional criterion helps maintain the dataset's robustness by ensuring that users have sufficient non-mental health content for analysis.

3.2 Control Matching

Each diagnosed user is paired with a set of control users who exhibit similar interests and posting behaviors, achieved by selecting control users from the same subreddits as the diagnosed users. To align posting habits, the number of posts made by each control user is ensured to fall within a specified range relative to their diagnosed counterpart's post count as described in (Coppersmith et al., 2018). Each diagnosed users is paired with 9 control users, as per (Coppersmith et al., 2018). Control users are subject to the same exclusion criteria as diagnosed users, but are excluded entirely in case they posted about mental health content, instead of excluding the mental health content. This additional criterion makes it challenging to find suitable control users for every diagnosed individual.

3.3 Descriptive Stats of Dataset

Table 1 shows statistics of the data we collected for seven conditions and the control group. As the table shows, our most common conditions are ADHD, Autism, and Bipolar.

Table 1 highlights that ADHD is the most common condition in our dataset. This skew is likely due to the high activity level and engagement

within ADHD-related subreddits compared to others. Users in ADHD communities often share frequent updates, experiences, and coping strategies, leading to a higher volume of posts and greater visibility of self-reported diagnoses. Additionally, ADHD has become a more widely discussed topic on social media platforms, attracting larger communities and increasing the likelihood of data collection from these users. This concentration may skew the dataset, requiring careful consideration during analysis to ensure that findings are not overly influenced by this imbalance.

4 Analysis

4.1 Non-English Content Analysis

Given the multilingual nature of Reddit, it was necessary to ensure that the dataset consisted mainly of English content. A random sample of 1,000 control posts was manually inspected for non-English text, revealing that 83 posts contained non-English content. All non-English posts were then automatically identified using the `langdetect` Python library[2] and were excluded from further analysis to maintain consistency and relevance in the study of English-language discourse. As Table 3 shows, the percentage of non-English content was relatively low for diagnosed users but higher for controls (3.02%). This discrepancy may be attributed to the broader and more diverse nature of control group discussions compared to diagnosed groups, which are more focused and specific. Ensuring linguistic uniformity is crucial, as even a small amount of non-English content can introduce noise, reducing the accuracy of natural language processing tasks like sentiment analysis and topic modeling (Baldwin and Lui, 2010).

4.2 Duplicate Content Removal (Within and Across Conditions)

During data preprocessing, we identified instances of duplicate content both within a single condition and across multiple conditions, often arising when users submitted identical or highly similar content more than once. We also found control users who were matched with more than one diagnosed user across different conditions.

Table 3 shows considerable variability in the percentage of duplicate posts across conditions. *Anxiety* has the highest duplication rate at **45.36%**, while *ADHD* also shows a high duplicate rate of **34.35%**. This suggests that users often post the same content multiple times across different subreddits, likely in an effort to receive more replies or engagement. On the other hand, *PTSD* displays the lowest duplicate percentage at **0.73%**, indicating fewer repeated posts among users discussing PTSD. Such patterns highlight the need for careful filtering of duplicates to maintain data quality.

These patterns help explain the results observed in the machine learning models, where *PTSD* achieved the highest accuracy and F1 score across classifiers (as seen in Table 4). The lower rate of duplicate content in *PTSD* posts likely contributes to a cleaner and more reliable dataset, improving model performance. Conversely, the high duplication rate in *ADHD* may introduce noise, leading to relatively lower accuracy and F1 scores in the ADHD classification models.

4.3 Feature Importance

To investigate the effectiveness of our data in discriminating between diagnosed and control users, we used the LIWC lexicon (Pennebaker et al., 2015). LIWC (Linguistic Inquiry and Word Count) analyzes text to provide quantitative measures of various linguistic and psycholinguistic categories, including emotional tone, cognitive processes, and social references. As seen in Table 2, the LIWC categories that showed the largest effect sizes across multiple conditions included *Negative Emotion* (e.g., anxiety and anger), *Cognitive Processes* (e.g., analytic thinking and insight), and *Pronouns* (e.g., first-person singular and third-person plural), highlighting the distinct linguistic patterns associated with different mental health conditions. These findings emphasize how individuals with specific conditions use language differently, providing valuable indicators for classification.

5 Experiments

To evaluate the effectiveness of LIWC features in distinguishing between different mental health conditions, we conducted a series of classification experiments using four machine learning models: Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting. The dataset was divided by condition, and each model was trained and evaluated separately for each condition using LIWC numerical data. The

[1] Coppersmith et al. (2018), SMHD: Social Media and Mental Health Dataset.

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

Table 2: Effect sizes (Cohen's d) of LIWC features comparing diagnosed and control groups. Blue shading indicates features more prevalent in the diagnosed group, while red shading indicates features more prevalent in the control group. The intensity of the shading reflects the strength of the feature in distinguishing between the two groups. Table structure inspired by the SMHD dataset methodology.

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
Apostro	0.17	0.13	0.35	0.09	0.24	-	0.19
Authentic	0.38	0.59	0.40	0.51	0.47	0.56	0.52
BigWords	0.09	-0.16	0.17	-0.1	-0.16	-0.14	-0.16
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
Family	0.05	0.06	0.56	0.03	0.04	0.08	0.07
Friends	0.10	0.09	0.31	0.08	0.10	0.12	0.14
Achievement	0.04	0.03	0.05	0.02	0.01	0.05	0.06
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	-0.03	-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
Joy	0.06	0.07	0.33	0.09	0.08	0.10	0.12
Confidence	0.02	0.03	0.30	0.01	0.00	0.04	0.03

Table 3: Percentage of Duplicated and Non-English Posts Identified from all Conditions. All duplicated and non English content is removed.

Condition	Dups (%)	Non-Eng (%)
Controls	12.21%	3.03%
ADHD	34.35%	1.01%
Autism	9.89%	1.29%
Anxiety	45.36%	0.66%
Bipolar	24.19%	0.96%
Depression	30.93%	1.39%
OCD	37.16%	1.12%
PTSD	0.73%	0.46%

performance metrics considered include Accuracy, Recall, Precision, and F1 Score.

The results of these experiments, as summarized in Table 4, show that the models performed consistently across the various conditions, demonstrating the robustness of the LIWC features. The accuracy for all models remained high, with PTSD and Bipolar conditions achieving the highest accuracy scores across all classifiers. SVM and Gradient Boosting showed slightly higher performance in terms of precision and F1 score compared to other models, especially for conditions like PTSD and Autism. The variability in performance across conditions indicates that certain LIWC features may be more informative for some conditions than others.

Additionally, the consistency in recall across conditions highlights the models' capability to correctly identify positive cases, emphasizing the effectiveness of the LIWC features in capturing language patterns associated with different mental health conditions.

6 Conclusion

This research aimed to construct a high-quality dataset for mental health analysis in the post-COVID era by applying a rigorous preprocessing pipeline. We filtered non-English content, removed duplicates, and enforced strict criteria to ensure the dataset accurately reflects diagnosed users' language patterns on Reddit. Using this dataset, we identified significant LIWC features, with effect sizes (Cohen's d) highlighting the distinct linguistic markers for conditions such as PTSD and Bipolar. Our machine learning models, including SVM and Gradient Boosting, achieved high accuracy and precision across conditions, demonstrating the effectiveness of our dataset for mental health analysis. This dataset serves as a valuable resource for understanding the evolving impact of the COVID-19 pandemic on mental health and supports developing effective early detection systems and large-scale studies.

7 Limitations

This study presents several limitations that must be acknowledged. First, the dataset used in this research may not fully capture the diversity of mental health conditions across different populations, as it primarily consists of social media data from a specific demographic e.g., male dominated. This limitation may affect the generalizability of the findings. Additionally, while the models employed demonstrate promising performance, there is room for improvement in handling nuanced and context-specific aspects of mental health, which were not extensively explored in this study.

8 Ethical Considerations

The use of social media data, such as the dataset employed in this research, raises important ethical considerations that must be addressed, particularly regarding consent and the responsible handling of user-generated content. While the data collected from Reddit is publicly available, it is crucial to acknowledge that users may not have explicitly consented to their posts being utilized for research purposes. This lack of informed consent necessitates a careful and respectful approach to the data, ensuring that the privacy and anonymity of users are upheld.

In future research utilizing this dataset, caution must be exercised to mitigate potential ethical concerns. Importantly, no attempt was made to contact users or reveal their anonymity throughout this study. Researchers should ensure that any analyses conducted do not compromise user anonymity or reveal identifiable information, as this could lead to unintended consequences for the individuals involved. Additionally, researchers are encouraged to consider the implications of their findings and how they may impact the communities represented in the dataset. By prioritizing ethical considerations in research practices, we can foster a more responsible and respectful use of social media data, ultimately benefiting both the scientific community and the individuals whose voices contribute to our understanding of mental health.

Table 4: Model Performance Metrics (Accuracy, Recall, Precision, F1) for Various Conditions

Condition	Logistic Regression	SVM	Random Forest	Gradient Boosting
ADHD	Acc: 0.83 Rec: 0.82 Prec: 0.84 F1: 0.83	Acc: 0.83 Rec: 0.82 Prec: 0.84 F1: 0.83	Acc: 0.84 Rec: 0.80 Prec: 0.87 F1: 0.83	Acc: 0.85 Rec: 0.84 Prec: 0.85 F1: 0.84
Autism	Acc: 0.86 Rec: 0.85 Prec: 0.86 F1: 0.86	Acc: 0.86 Rec: 0.85 Prec: 0.87 F1: 0.86	Acc: 0.85 Rec: 0.83 Prec: 0.87 F1: 0.85	Acc: 0.86 Rec: 0.86 Prec: 0.87 F1: 0.86
Bipolar	Acc: 0.88 Rec: 0.86 Prec: 0.89 F1: 0.88	Acc: 0.88 Rec: 0.87 Prec: 0.89 F1: 0.88	Acc: 0.88 Rec: 0.85 Prec: 0.90 F1: 0.88	Acc: 0.89 Rec: 0.87 Prec: 0.90 F1: 0.89
Depression	Acc: 0.86 Rec: 0.84 Prec: 0.87 F1: 0.86	Acc: 0.86 Rec: 0.84 Prec: 0.87 F1: 0.86	Acc: 0.84 Rec: 0.81 Prec: 0.86 F1: 0.84	Acc: 0.86 Rec: 0.84 Prec: 0.87 F1: 0.85
OCD	Acc: 0.86 Rec: 0.86 Prec: 0.87 F1: 0.87	Acc: 0.86 Rec: 0.86 Prec: 0.87 F1: 0.87	Acc: 0.85 Rec: 0.83 Prec: 0.87 F1: 0.85	Acc: 0.87 Rec: 0.87 Prec: 0.89 F1: 0.88
PTSD	Acc: 0.90 Rec: 0.91 Prec: 0.91 F1: 0.91	Acc: 0.90 Rec: 0.90 Prec: 0.91 F1: 0.90	Acc: 0.89 Rec: 0.87 Prec: 0.91 F1: 0.89	Acc: 0.90 Rec: 0.90 Prec: 0.90 F1: 0.90
Anxiety	Acc: 0.86 Rec: 0.84 Prec: 0.88 F1: 0.86	Acc: 0.86 Rec: 0.84 Prec: 0.88 F1: 0.86	Acc: 0.86 Rec: 0.81 Prec: 0.90 F1: 0.85	Acc: 0.87 Rec: 0.84 Prec: 0.89 F1: 0.87

References

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

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.

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.

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.

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

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. 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.