# Leveraging LLMs for Mental Health: Detection and Recommendations from Social Discussions

Vaishali Aggarwal
*TCS Research*
Delhi, India
aggarwal.vaishali@tcs.com

Sachin Thukral
*TCS Research*
Delhi, India
sachi.2@tcs.com

Krushil Patel
*IIT Kharagpur*
West Bengal, India
krushilpatel@kgpian.iitkgp.ac.in

Arnab Chatterjee
*TCS Research*
Delhi, India
arnab.chatterjee4@tcs.com

***Abstract*—Textual data from social platforms captures various aspects of mental health through discussions around and across issues, while users reach out for help and others sympathize and offer support. We propose a comprehensive framework that leverages Natural Language Processing (NLP) and Generative AI techniques to identify and assess mental health disorders, detect their severity, and create recommendations for behavior change and therapeutic interventions based on users' posts on Reddit.**

**To classify the disorders, we use rule-based labeling methods as well as advanced pre-trained NLP models to extract nuanced semantic features from the data. We fine-tune domain-adapted and generic pre-trained NLP models based on predictions from specialized Large Language Models (LLMs) to improve classification accuracy. Our hybrid approach combines the generalization capabilities of pre-trained models with the domain-specific insights captured by LLMs, providing an improved understanding of mental health discourse. Our findings highlight the strengths and limitations of each model, offering valuable insights into their practical applicability.**

**This research potentially facilitates early detection and personalized care to aid practitioners and aims to facilitate timely interventions and improve overall well-being, thereby contributing to the broader field of mental health surveillance and digital health analytics.**

***Index Terms*—Mental health, behavior, Reddit, GenAI**

## I. Introduction

Mental health is crucial for overall well-being, with early detection and intervention being vital. Mental illness affects millions worldwide and is a significant health burden, worsened by the COVID-19 pandemic [1]. Social media, especially Reddit, offers valuable insights into mental health through extensive user discussions around experiences and struggles. However, analyzing this data is challenging [2] due to informal language, slang, and emojis, which complicate traditional NLP techniques.

Generative AI, particularly Large Language Models (LLMs), has advanced the analysis of unstructured mental health data but poses risks like misinformation and misdiagnosis. Ethical concerns, biases, and the need for human oversight are crucial. Ongoing efforts aim to improve LLMs while addressing these challenges.

Advanced pre-trained NLP models are essential for processing text – they excel in classification, sentiment analysis, and semantic understanding. Learning from labeled data, they improve classification performance and effectively interpret nuanced language in social media discussions.

We propose a framework to leverage the knowledge of LLMs to: (i) **Identify Mental Health Issues**: Utilize social media data to detect whether it is related to mental health, (ii) **Classify Disorders and Severity**: Accurately classify the disorders mentioned in the posts and assess their severity, and (iii) **Recommend Interventions**: Provide recommendations for behavior changes and therapies based on the identified disorders and their severity.

This paper presents a framework that utilizes large-scale Reddit data on mental health, combining LLMs with advanced NLP models for a multi-label classification. This integration aims to enhance the precision and reliability of mental health analysis by addressing the complexities of social media text and bridging the gap between vast social media data and actionable mental health insights.

The research has the potential to advance mental health informatics by enabling more accurate identification of mental health disorders from social media, potentially facilitating earlier interventions and improving lives.

## II. Related work

Analyzing social media data for mental health diagnosis has become increasingly significant due to the wealth of information available on social media platforms like Twitter, Reddit, etc. Various methodologies developed to enhance the accuracy of identifying mental health conditions from user-generated content. Linguistic markers of depression examined from Twitter data [3] highlighted the potential of social media for mental health insights. Recent work on identifying stress factors that lead to mental health decline has demonstrated the effectiveness of NLP techniques in this domain [4]. Various machine learning methods, particularly deep learning, have improved detection accuracy for mental health issues on social media [5]. The "Dreaddit" dataset showcased NLP techniques for identifying stress-related content on Reddit through binary classification [6]. Additionally, a gold standard dataset [7] was developed for detecting depression levels using traditional algorithms and data augmentation in a multi-class framework. RoBERTa [8] enhances BERT's performance through optimized pre-training, but our focus on mental health aligns better with domain-specific models like MentalRoBERTa [9]. "Illuminate" [10] leverage GPT-4 and Llama2 for depression detection and therapy recommendations through advanced prompt

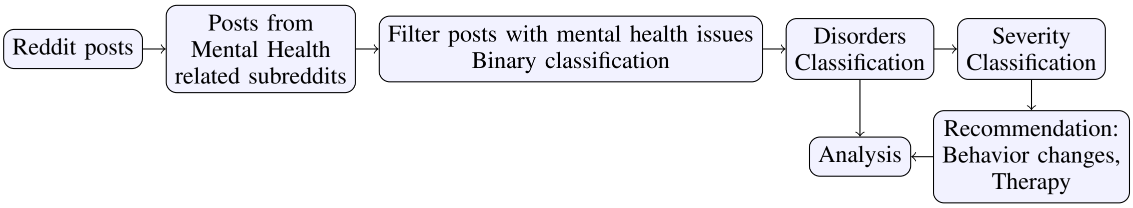
Reddit posts
Posts from Mental Health related subreddits
Filter posts with mental health issues Binary classification
Disorders Classification
Severity Classification
Analysis
Recommendation: Behavior changes, Therapy

Fig. 1. Framework for analyzing Reddit posts on mental health, with modules for disorders classification, severity detection, behavior change recommendation and therapy recommendation.

engineering. A study analyzing LLMs as therapists [11] suggests they exhibit low-quality therapy behaviors but also address client needs, indicating the need for further research to improve care quality.

## III. Dataset

Reddit is a popular platform with active communities organized into subreddits. We extracted text content from posts in the subreddits `r/mentalhealth` and `r/depression` for the period of December 1, 2019 to November 30, 2020, totaling 334,197 posts using the Pushshift [12] API. Each post is treated as an independent entity, regardless of the user, to focus on content analysis. To validate our methodology, we selected a representative subset of 5,000 random posts from the dataset, ensuring it captures the essential characteristics for evaluating our approach.

## IV. Methodology

Our proposed framework, outlined in Fig. 1 for identifying mental health conditions from Reddit posts, involved several key steps, which are elaborated below.

*Leveraging Large Language Models (LLMs)*

We used Llama3 [13], MentaLLaMA [14], and Samantha-Mistral for our analysis, exploring their potential for mental health condition identification and analysis. MentaLLaMA, based on LLaMA2 models, offers significant improvements in providing both accurate classifications and high-quality explanations for mental health analysis on social media. MentaLLaMA addresses the challenge of low interpretability in traditional mental health analysis methods by building the Interpretable Mental Health Instruction (IMHI) dataset. Samantha-Mistral LLM, also known as Samantha-1.2-Mistral-7B[1], is an advanced language model developed to function as a caring and empathetic AI companion, and has been trained extensively in philosophy, psychology, and personal relationships, making it uniquely suited to engage in human-like interactions and offer support in these areas.

### *A. Preprocessing using Binary Classification*

We employed various LLMs to filter out irrelevant content, thereby refining the dataset to include only posts that were genuinely related to mental health conditions. Each post was evaluated to determine whether it pertained to mental health, with the LLMs instructed to provide responses strictly in *yes* or *no* format. By leveraging the advanced capabilities of LLMs, we were able to efficiently and accurately classify the vast number of posts, ensuring the integrity and focus of our analysis on meaningful mental health-related content (see Table. I).

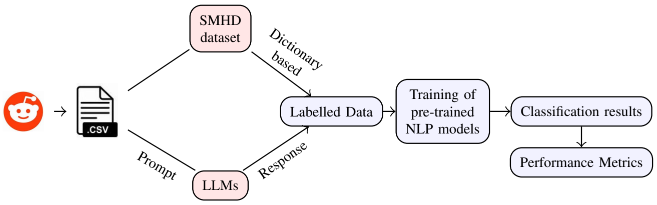

Fig. 2. Framework for disorders classification for Reddit posts on mental health.

TABLE I
Preprocessing for mental health related posts

| Model Name | Yes | No | Other |
|---|---|---|---|
| Llama3 | 67.7% | 27% | 5.3% |
| MentaLLaMA | 85.9% | 12% | 2.1% |
| Samantha-Mistral | 58.6% | 41.3% | 0% |

### *B. Disorders Classification*

*1) Data Labeling:* For this study, we selected nine labels based on the Self-reported Mental Health Diagnoses (SMHD) dataset [15] which contains dictionaries for nine distinct mental health disorders – *Attention-Deficit Hyperactivity Disorder (ADHD), Autism, Anxiety, Bipolar Disorder, Depression, Eating Disorder, Obsessive-Compulsive Disorder (OCD), Post-Traumatic Stress Disorder (PTSD)* and *Schizophrenia*. This selection allows us to comprehensively cover a range of mental health issues within our analysis.

*2) Data Annotation:* We employed four different annotators to label Reddit posts for mental health disorders. These annotators included (i) dictionary-based labeling, (ii) Llama3, (iii) Samantha-Mistral, and (iv) MentaLLaMA.

We used SMHD dataset dictionaries to label Reddit posts by matching terms linked to mental health disorders. Posts with terms from multiple dictionaries received multiple labels, while unmatched posts were labeled as *None*. For the LLM based annotations, a suitable prompt classified posts with pre-

[1] https://huggingface.co/cognitivecomputations/samantha-1.2-mistral-7b

defined labels for different disorders. The overall framework of disorders classification is shown in Fig. 2.

To ensure consistency, we filtered predictions to include only labels for the nine predefined mental health disorders. This multi-annotator approach allowed us to effectively compare traditional dictionary-based methods with advanced LLMs in identifying mental health disorders.

*3) Training Pre-trained Models:* We evaluated various pre-trained NLP models, including domain-specific ones like MentalRoBERTa [9], and general models like DistilBERT [16]. Each model was fine-tuned and trained to classify posts based on labeled data from different annotators to improve accuracy. Input sequences were tokenized using respective pre-trained tokenizers, truncated, and padded to a maximum of 512 tokens. Both models were fine-tuned using the Binary Cross-Entropy loss function with the AdamW optimizer, optimizing them for identifying mental health disorders in Reddit posts.

*4) Model Evaluation:* We evaluated the performance of both pre-trained NLP models and LLMs using standard metrics such as accuracy, precision, recall, and F1-score. For validation, a 5-fold cross-validation strategy was employed, training each fold for 10 epochs. We assessed the generalizability of our trained models on a new, unseen sample of Reddit posts related to mental health, comparing their predictions with dictionary-assigned labels and LLM predictions. This comprehensive evaluation framework aimed to improve the accuracy and robustness of mental health disorders identification, facilitating timely interventions and better well-being outcomes.

### C. Severity Classification

We utilized various LLMs to assess the severity level of each user's post, classifying them into three distinct categories: *mild*, *moderate*, and *severe*.

### D. Recommendation Module

*1) Suitable Therapy:* This module leverages LLMs to provide personalized therapy suggestions based on the severity of each user post. LLMs identify suitable therapies and offer a confidence percentage for each, allowing for comparison and ensuring the recommendations are both appropriate and backed by a quantifiable measure.

*2) Behaviour Change:* The recommendation module also includes suggestions for behavior changes to improve mental well-being. For each user post, we generated tailored recommendations for behavior changes that could positively impact the user's mental health, crafted based on the specific context and content of the posts, ensuring that they are relevant and actionable.

The above provides personalized behavior advice practical strategies to help users improve mental well-being, complementing therapy for a holistic mental health approach.

## V. Results

In the rest of the paper, we demonstrate our results on a random sample of 5000 posts.

### A. Preprocessing using Binary Classification

The results corresponding to binary classification task for various LLMs are given in Table. I. Here, *Other* denotes the cases where the LLM failed to perform the classification. We observed that MentaLLaMA, being a domain-specific model, classifies the majority of the data as being related to mental health disorders. The set of posts that are classified as "YES" in the above constitutes our final dataset for further analysis.

### B. Disorders Classification

During fine-tuning, we experimented with different learning rates for both DistilBERT and MentalRoBERTa, achieving the best accuracy with $3e-5$.

Table II presents the averaged results across folds, obtained after excluding 'None' labels from each annotator.

As shown in Table II, both DistilBERT and MentalRoBERTa demonstrated varying performance across different annotators. Llama3 and MentaLLaMA achieved the highest training accuracy (95.85% and 94.15% for DistilBERT, 91.02% and 92.35% for MentalRoBERTa), but exhibited significant over-fitting or sensitivity to some specific disorders it was trained on, as evidenced by the substantial drop in validation accuracy (70.31% and 59.21% for DistilBERT, 69.42% and 60.64% for MentalRoBERTa). Dictionary matching offered consistent and stable performance, particularly in terms of validation accuracy (86.25% with DistilBERT and 82.73% with MentalRoBERTa) but was limited by its reliance on explicit keyword matching. Fig. 3 highlights the class imbalance in the dictionary matching dataset, with a high concentration of '*None*' labels (50.5%), followed by *Depression* (30.26%) and *Anxiety* (10.93%), which might have contributed to its relatively strong performance due to potential biases towards majority classes. However, this approach is inherently limited in capturing subtle emotional cues and contextual details, which constrains its ability to fully comprehend the nuances of the posts. The F1 Scores for Llama3 (0.88 for both) and dictionary Matching (0.87 for MentalRoBERTa and 0.85 for DistilBERT) were the highest among the annotators, suggesting that these models performed better in maintaining a balance between precision and recall. While dictionary matching offers consistency and accuracy, Llama3 and MentaLLaMA demonstrated potential for capturing nuanced language patterns, although with a higher risk of over-fitting. Samantha-Mistral, despite its overall low performance, provided valuable insights into areas requiring further model refinement. Overall, the DistilBERT model, when fine-tuned with Llama3, provided the best generalization across the different annotators.

Limity

*Analysis of labels:* The distribution of different labels across annotators is shown in Fig. 3 while the distribution of label counts per post for each annotator is presented in Fig. 4.

From Fig. 3, we observe that *Depression* and *Anxiety* were the most frequently assigned labels across all annotators, with a pronounced dominance in the dictionary-based method. This suggests these disorders are either highly prevalent or more easily detectable using keyword-based approaches. Llama3

TABLE II
NLP Model Evaluation

| | Models | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Annotators | DistilBERT | | | | | Mental RoBERTa | | | | |
| | Tr. Acc. % | Val. Acc. % | Precision | Recall | F1 Score | Tr. Acc. % | Val. Acc. % | Precision | Recall | F1 Score |
| Dictionary Matching | 87.7 | 86.25 | 0.88 | 0.87 | 0.87 | 84.16 | 82.73 | 0.83 | 0.87 | 0.85 |
| Llama3 | 95.8 | 70.31 | 0.88 | 0.88 | 0.88 | 91.02 | 69.42 | 0.87 | 0.91 | 0.88 |
| MentaLLaMA | 94.15 | 59.21 | 0.81 | 0.85 | 0.82 | 92.34 | 60.64 | 0.78 | 0.8 | 0.79 |
| Samantha-Mistral | 83.47 | 32.40 | 0.71 | 0.82 | 0.75 | 79.26 | 35.24 | 0.76 | 0.80 | 0.76 |

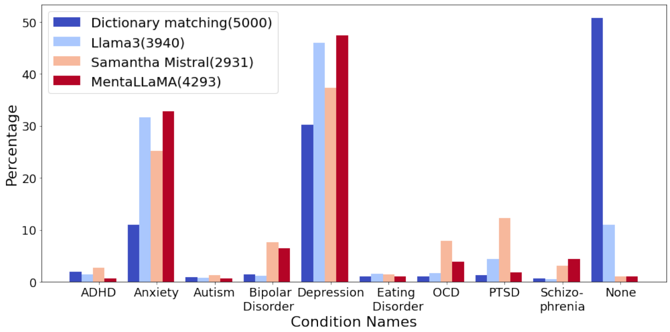

Fig. 3. Normalized condition counts for various annotators

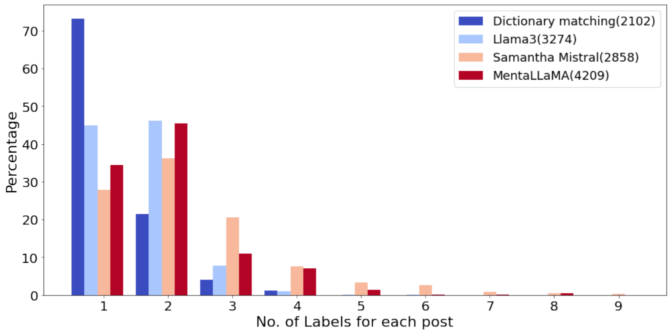

Fig. 4. Number of labels for each post for various annotators

and MentaLLaMA demonstrated a more balanced distribution of labels, indicating better detection of less common disorders. A substantial portion of posts were labeled as '*None*' by the dictionary-based method (50.74%), implying either a conservative labeling approach or the absence of relevant keywords in the text whereas for Llama3 it was 10.93%, which consisted mostly of the suicidal/self-harm cases. For MentaLLaMA and Samantha Mistral, it turned out to be less than 1%.

Fig. 4 shows that most posts receive a single label for dictionary-based annotation which turned out to be more conservative. In contrast, Large Language Models (LLMs) tended to assign multiple labels, suggesting a greater capacity to identify overlapping disorders. Across all annotators, the frequency of posts decreased with increasing label count. While the dictionary-based method identified up to six labels per post, LLMs detected up to nine, indicating their enhanced ability to recognize multiple disorders.

### C. Severity Classification

Following the pre-processing step, we used various LLMs to categorize the severity level of each post into three classes – *mild*, *moderate*, and *severe*, as described in Sec. IV-C. All three LLMs classified the majority of the posts into the *Severe* category, with Samantha-Mistral classifying no posts as *Mild* (see Table. III). The LLMs identified and interpreted indicators of distress, disorder severity, and urgency in mental health discussions, enabling the categorization of posts by disorder intensity. This allows for targeted analysis and appropriate therapeutic recommendations, making it a key component in understanding and addressing varying mental health challenges.

TABLE III
Analysis of Severity Level Classification

| **Model Name** | **Mild** | **Moderate** | **Severe** |
|---|---|---|---|
| Llam3 | 1.5% | 17.4% | 81.1% |
| MentaLLaMA | 7.4% | 10.3% | 82.3% |
| Samantha-Mistral | 0% | 3% | 97% |

### D. Recommendation Module

Our queries on LLMs produced a list of recommended therapies. Highly recommended therapies include Cognitive Behavioral Therapy (CBT), Dialectical Behavior Therapy (DBT), Acceptance and Commitment Therapy (ACT), Psychodynamic Therapy (PT), Mindfulness-Based Stress Reduction (MBSR), Mindfulness-Based Cognitive Therapy (MBCT), Interpersonal Therapy (IPT), Exposure Therapy (ET), Motivational Interviewing (MI), and Family Therapy (FT). Other Therapy (OT) groups all the less frequent therapies, while No Therapy (NT) is suggested when LLMs recommend users to seek professional help, particularly in 'extreme' cases where it detects a tendency towards self-harm or suicidal thoughts.

CBT is the most recommended therapy for all three LLMs as a result of its adaptability and broad applicability. However, their second most recommended differ: Llama3 prefers ACT, MentaLLaMA prefers DBT, whereas Samantha-Mistral focuses on less frequent therapies.

We mapped the disorders detected in posts to the therapies recommended in them. We observed that ACT is more recommended than CBT by Llama3 for disorders like anxiety and depression. This finding is significant, as CBT has traditionally been considered the standard approach for these problems. However, Llama3's preference for ACT suggests a greater emphasis on mindfulness, acceptance, and commitment to personal values, highlighting the model's inclination towards therapies that focus on holistic mental well-being.

We have also asked the LLMs to recommend behavior changes for these posts (see Sec. IV-D). We observe that LLMs are suggesting actionable tasks to the user to improve their mental health. The LLMs provide behavior change suggestions aimed at practical daily actions for users to incorporate into their lives. These include mindfulness techniques, healthy routines, and coping mechanisms for mental health issues. The suggestions focus on small, manageable tasks to foster a positive outlook and empower individuals to take control of their mental well-being. This approach addresses immediate emotional needs and supports long-term mental health, laying a foundation for sustained personal growth and self-care.

## VI. Time complexity of different LLMs

The evaluation of three LLMs – Llama3, MentalLaMA, and Samantha-Mistral – was conducted on a TP100 GPU with 15 GB memory. Llama3, the largest and most parameter-heavy model, took 40 minutes to process 100 samples. MentaLLaMA completed the same task in 19 minutes, while Samantha-Mistral took 100 seconds. However, Samantha-Mistral's speed comes with reduced effectiveness, making it less suitable for complex tasks. In summary, Llama3 excels in capability but demands resources, MentalLaMA balances speed and accuracy, and Samantha-Mistral, though fast, struggles with deeper analysis.

## VII. Discussions

In this paper, we present the comprehensive framework for the evaluation of multiple pre-trained NLP models and LLMs involving binary, multi-class, multi-label classification, and recommendation tasks from social media discussions related to mental health. We observed that while the dictionary-matching approach is more stable, it lacked the depth required to identify disorders based on the context and sentiment of the posts, which could limit its effectiveness in complex scenarios. This can be further fine-tuned by including stylistic features, emotion, and Empath libraries to make it more reliable. Our proposed framework is not only robust in identifying a wide range of mental health disorders but also versatile enough to provide tailored recommendations for therapy and behavior changes. Although we present results for a sample dataset, the framework can be scaled up for much larger datasets, making it a valuable tool for large-scale mental health assessment and recommendations.

Moreover, careful prompt engineering of LLMs has proven effective in generating comprehensive insights, including disorder identification, severity assessment, recommendation of therapy and behavior changes. We observed that Llama3 outperformed the other two LLMs in disorder identification. For other tasks, MentaLLaMA delivered better results. However, Samantha-Mistral, although being the fastest, has not yet met the performance standards for these tasks.

Ethical considerations are essential throughout this process. While the framework prioritizes accurate analysis, language models are designed to handle sensitive content with care, often refusing to generate output related to suicide or self-harm to uphold ethical standards and ensure user safety.

Our framework has the potential to contribute significantly to the field of mental health informatics. By developing more accurate methods for automatically identifying mental health disorders from social media data, we can potentially enable earlier interventions and improve the lives of individuals struggling with these problems. These insights could be instrumental for mental health practitioners, offering them a new resource for monitoring and intervening in patient care. Additionally, the methodology has the capability to supplement a digital well-being platform, providing personalized mental health support directly to users.

## References

[1] D. F. Santomauro *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[2] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 71–80.

[3] J. Wolohan *et al.*, "Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP," in *Proceedings of the first international workshop on language cognition and computational models*, 2018, pp. 11–21.

[4] S. Thukral *et al.*, "Identifying pandemic-related stress factors from social-media posts–Effects on students and young-adults," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[5] M. M. Tadesse *et al.*, "Detection of depression-related posts in Reddit social media forum," *Ieee Access*, vol. 7, pp. 44 883–44 893, 2019.

[6] E. Turcan and K. McKeown, "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media," *EMNLP-IJCNLP 2019*, p. 97, 2019.

[7] K. Sampath and T. Durairaj, "Data set creation and empirical analysis for detecting signs of depression from social media postings," in *International Conference on Computational Intelligence in Data Science*. Springer, 2022, pp. 136–151.

[8] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, E. Holderness *et al.*, Eds. online: Association for Computational Linguistics, Apr. 2021.

[9] S. Ji *et al.*, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 2022, pp. 7184–7190.

[10] A. Agrawal, "Illuminate: A novel approach for depression detection with explainable analysis and proactive therapy using prompt engineering," *arXiv preprint arXiv:2402.05127*, 2024.

[11] Y. Y. Chiu *et al.*, "A computational framework for behavioral assessment of llm therapists," *arXiv preprint arXiv:2401.00820*, 2024.

[12] Pushshift, "Calling API for dumping dataset," https://pushshift.io/api-parameters, 2021, accessed: 2021-01-01.

[13] A. Dubey *et al.*, "The Llama 3 Herd of Models," 2024.

[14] K. Yang *et al.*, "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models," in *Proceedings of the ACM on Web Conference 2024*, ser. WWW '24. ACM, May 2024.

[15] A. Cohan *et al.*, "SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions," in *27th International Conference on Computational Linguistics*. ACL, 2018, pp. 1485–1497.

[16] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.