Depression Detection on Social Media with Large Language Models

Xiaochong Lan, Yiming Cheng, Li Sheng, Chen Gao, Yong Li
Department of Electronic Engineering, BNRist, Tsinghua University, China
lanxc22@mails.tsinghua.edu.cn, {chgao96, liyong07}@tsinghua.edu.cn

Abstract

Depression harms. However, due to a lack of mental health awareness and fear of stigma, many patients do not actively seek diagnosis and treatment, leading to detrimental outcomes. Depression detection aims to determine whether an individual suffers from depression by analyzing their history of posts on social media, which can significantly aid in early detection and intervention. It mainly faces two key challenges: **1)** it requires professional medical knowledge, and **2)** it necessitates both high accuracy and explainability. To address it, we propose a novel depression detection system called DORIS, combining medical knowledge and the recent advances in large language models (LLMs). Specifically, to tackle the first challenge, we proposed an LLM-based solution to first annotate whether high-risk texts meet medical diagnostic criteria. Further, we retrieve texts with high emotional intensity and summarize critical information from the historical mood records of users, so-called *mood courses*. To tackle the second challenge, we combine LLM and traditional classifiers to integrate medical knowledge-guided features, for which the model can also explain its prediction results, achieving both high accuracy and explainability. Extensive experimental results on benchmarking datasets show that, compared to the current best baseline, our approach improves by 0.036 in AUPRC, which can be considered significant, demonstrating the effectiveness of our approach and its high value as an NLP application.

1 Introduction

Depression profoundly affects humanity, with WHO estimates indicating that 5% of adults suffer from it[1], significantly contributing to the global suicide rate[2]. Given the harmful nature of depression, timely diagnosis and intervention are necessary. However, traditional hospital-based approaches for diagnosing depression face several issues. Firstly, patients often avoid evaluations due to stigma or not recognizing their need for help [34]. What's more, self-reported diagnoses can be unreliable due to intentional concealment [4]. Additionally, the high cost of hospital evaluations places a burden on patients [22]. For these reasons, many individuals with depression remain undiagnosed and untreated, with over 75% of those in low- and middle-income countries receiving no treatment at all[3].

Depression detection on social media identifies potential depression through users' post histories [29, 35], as shown in Figure 1, which is one typical beneficial application of NLP techniques. That is, it leverages public posts from online social networks for broad detection coverage, benefits from more genuine expressions than those in clinical settings [23], and reduces economic costs compared to professional diagnoses.

[1]www.who.int/news-room/fact-sheets/detail/depression
[2]www.ncbi.nlm.nih.gov/pmc/articles/PMC6165520/
[3]press.un.org/en/2021/sgsm20951.doc.htm

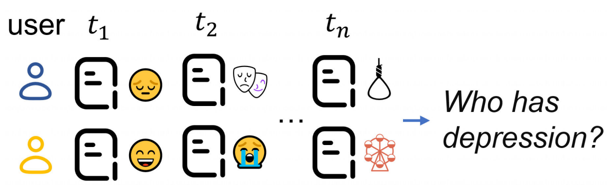


Figure 1: Illustration of depression detection on social media. It aims to determine whether a user has depression based on their history of posts on social media.

Many works have delved into depression detection based on texts from social media. [16] first identified the possibility of predicting medically diagnosed depression through expressions in social media posts. Recent works [24, 29, 42, 9] utilize deep learning methods to determine whether users suffer from depression based on their historical posts. However, they are not able to incorporate medical knowledge on depression, leaving significant room for improvement in both accuracy and interpretability. With the advent of LLMs, some studies have employed them for mental health analysis, achieving good explainability [41]. However, these methods fall short in accuracy [2], and are highly sensitive to small prompt variations [18].

To summarize, previous work has not resolved two significant challenges in depression detection as follows.

- First, depression detection necessitates reference to professional medical knowledge. Extensive research in medicine has led to widely used diagnostic methods and standards, including depression scales and mood course analysis. Integrating these explicit standards into automated systems presents a challenge, moving beyond the common use of black-box approaches like embedding models. It requires a nuanced analysis of posts and targeted modeling strategies.
- Second, depression detection demands both high accuracy and high explainability. As a serious and safety-critical task, it requires trustworthy methods with high performance in both two aspects. While traditional non-LLM methods often achieve higher accuracy, they typically lack explainability. In contrast, LLMs naturally explain their judgments but tend to be less accurate. Thus, designing an approach that combines their strengths, ensuring accuracy and explainability, is essential yet very challenging.

To address these challenges, in this work, we propose a depression-detection system named DORIS (short for **D**iagn**O**stic **CR**iteria-Guided Mood **HIS**tory-Aware). To tackle the first challenge, we incorporate the widely used depression scale DSM-V. Leveraging the strong text understanding capability, we adopt an LLM to assess whether each user's high-risk texts reflect symptoms of depression, thereby constructing expert features for each user. Moreover, to model users' mood courses, we use a text embedding model to filter posts with high emotional intensity and summarize these texts into users' mood courses using an LLM. These high-emotion texts and mood course descriptions are then vectorized to generate representations of mood course. To tackle the second challenge, we combine features from different spaces to produce our method's final judgment by using the GBT classifier, achieving high accuracy. Additionally, using the annotations of high-risk texts and descriptions of mood history generated during the model's operation, we create a comprehensive explanation of the system's judgment, achieving high explainability.

To validate the effectiveness of our method, we conduct extensive experiments, and our method shows an improvement of 0.036 over the current best baseline on the AUPRC metric. The ablation study demonstrates the effectiveness of each component of our method. The case study illustrates that our method can generate explainable diagnostic reports, achieving high explainability.

Our contribution can be summarized as follows:

(1) We are among the first to combine professional medical knowledge and advanced large language models in depression detection.

(2) We propose the depression detection system DORIS, aiming to integrate professional medical knowledge with advanced NLP techniques to generate accurate and interpretable judgments.

(3) We conduct extensive experiments demonstrating the accuracy of DORIS, the effectiveness of each module's design, and the high quality of the model's output explanations.

The organization of our paper is as follows. In Section 2, we review related work. Then, in Section 3, we clearly describe the problem definition and our method design. Afterward, in Section 4, we present our experimental results. Finally, in Section 5, we summarize our work and propose future directions.

2 Related Works

2.1 Depression Detection on Social Media

Compared to traditional medical diagnostic methods for depression conducted in hospitals, depression detection on social media has the advantages of lower concealment potential, wider coverage, and lower cost [26]. Early research works first extract text features, then apply traditional machine learning methods for classification. Common feature extraction methods in depression detection research include LIWC [38], TF-IDF, LDA [3], etc.; classifiers include SVM [37], Logistic Regression [12], Random Forest [6], among others. In recent years, with the development of deep learning technologies, deep learning techniques have been used in depression detection, such as CNN [11], RNN [19], PLM [30, 20], GCN [27]. These efforts have improved the accuracy of depression detection, advancing the field's research.

Some works on depression detection focus solely on analyzing individual texts to determine the presence of depressive symptoms [42, 9], the type of symptoms displayed [46], and the level of depression [47]. However, analyzing a user's post history is more informative, as it is common for individuals not suffering from depression to also occasionally post texts that exhibit depressive symptoms. Recent approaches to depression detection based on a user's post history, such as those by [16], [29], [24], [9], and [42], lack comprehensive integration of medical knowledge, limiting their accuracy and interpretability. Our research stands out as one of the initial efforts to thoroughly apply medical insights, leading to improved accuracy. Additionally, we are pioneers in employing LLMs for crafting systematic explanations of predictions, thereby increasing interpretability.

2.2 Large Language Models for Mental Health Analysis

Large language models (LLMs) have shown great potential in clinical applications due to its abundant prior knowledge and strong language generalization ability [36]. Recent works have introduced LLMs into mental health analysis. [2] presents a solid evaluation on the performance of Llama-2 [39] and ChatGPT[4] in mental health detection tasks, unveiling prospects along with challenges of LLM-based methods in mental health analysis. [41] further discover the applications of LLMs in both mental health detection tasks and reasoning tasks, which highlight LLM's excellent interpretability. [43] propose the first open-source LLM series for mental health analysis based on Llama-2 and fine-tuning techniques, and greatly enhances the accuracy and explanation quality compared with general-purpose LLMs. However, the challenge of LLM-based methods being weaker in terms of prediction accuracy compared to embedding models, especially when targeting specific downstream tasks, remains unresolved [2].

To our knowledge, our system is the first to be specifically designed for depression detection in the context of LLM for mental health analysis. Furthermore, we have facilitated a collaboration between LLMs and embedding models, attaining both high accuracy and interpretability, addressing deficiencies in previous designs.

3 Methodology

3.1 Problem Formulation

In this study, we aim to assess the risk of depression in individuals based on their post history. For a user u, we can collect their historical posts, denoted as $P = \{p_1, p_2, ..., p_n\}$. Each of the posts has a corresponding timestamp, denoted as $t_1, t_2, ..., t_n$. Our task is formalized as determining the depression label y for user u, given their history of posts P, where $y \in 0, 1$.

[4]https://openai.com/blog/chatgpt

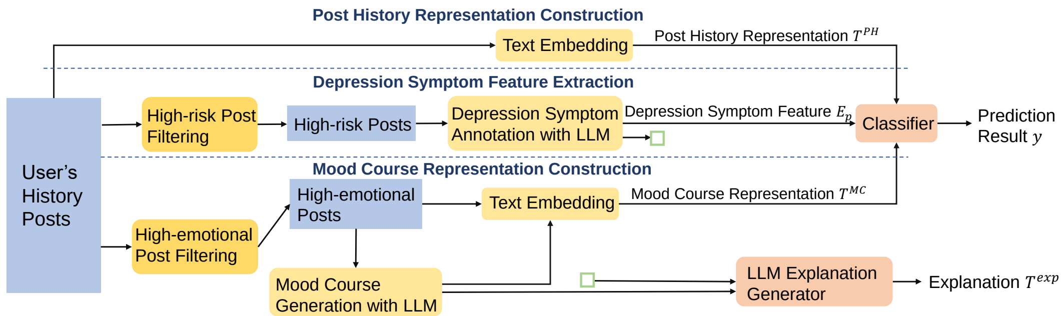


Figure 2: Illustration of DORIS. Through the collaboration of the LLM and the text embedding model, we obtain three key features: depression symptom feature, post history representation, and mood course representation. The classifier uses these three features to make its judgment; the LLM uses annotations of depression symptoms and descriptions of the mood course to generate explanation for the system's decision.

A. Depressed mood
B. Loss of interest/pleasure
C. Weight loss or gain
D. Insomnia or hypersomnia
E. Psychomotor agitation or retardation
F. Fatigue
G. Inappropriate guilt
H. Decreased concentration
I. Thoughts of suicide

Table 1: A concise summary of the symptoms of depression as defined in the DSM-5. To be diagnosed with depression, five (or more) of the listed symptoms must be present during the same two-week period.

3.2 Model Design

In this part, we describe our proposed DORIS in detail. The architecture of DORIS is illustrated in Figure 2.

3.2.1 Diagnostic Criteria Feature Extraction

Annotation with LLM In the medical domain, there are well-established diagnostic frameworks for depression, yet its full potential has not been tapped into for depression detection on social media. Our system seeks to incorporate these diagnostic criteria, enabling decisions supported by medical knowledge, thus improving decision accuracy and credibility. Specifically, we build user features based on medical diagnostic criteria,

We utilize the DSM-5 [32], a widely recognized tool for mental disorder assessment and diagnosis. It offers a list of clinical criteria for depression, detailed in Table 1.

LLMs exhibit a remarkable capacity for semantic understanding [1], and research has demonstrated their potential to replace human annotators in certain tasks [48]. Here, we employ LLMs to annotate texts, specifically to identify if and which self-expressed symptoms of depression are present in posts. The prompt is structured as follows:

Assuming you are a psychiatrist specializing in depression. Given 'text', please determine if this message includes any of the following states of the author:

A. Depressive mood B. Loss of interest/pleasure ... I. Thoughts of suicide.

If present, answer in the format of enclosed letters separated by commas, for example, (A, B, C). If none are present, respond with None.

By post-processing the output from the LLM, we can generate a 9-dimensional vector for each post, with each element being 0 or 1, indicating the absence or presence of a specific depression symptom. For instance, if the output for a post p is (G, I), then the corresponding vector E_p would be $(0, 0, 0, 0, 0, 0, 1, 0, 1)$.

Efficient Annotation The use of LLMs incurs substantial financial and energy costs. Since the vast majority of posts on social media platforms are unrelated to depression, annotating all posts would be exceedingly wasteful. To address this issue, we design an efficient annotation approach that first filters for high-risk texts and then annotates only those texts. Our experiments will demonstrate that annotating just the selected high-risk texts achieves a performance comparable to annotating all texts, while significantly reducing the number of requests to the LLM. Specifically, for each symptom, we designed a symptom template containing potential first-person textual expressions of that symptom. For example, for the symptom *B. Loss of interest/pleasure*, the template T_B^{DC} is:

I have lost interest, feel indifferent, bored, unconcerned, lack enthusiasm, am unmotivated, have no interest in activities, am unmotivated, find almost everything uninteresting, lack motivation, find significantly reduced pleasure, cannot experience happiness, feel the world is dull, and cannot muster energy all day.

Then, we get the text embedding of all symptom templates using a text embedding model:

$$H_i = \text{Encoder}(T_i^{DC}), \text{for } i = \text{A to I}. \tag{1}$$

For each post p, we compute its embedding as :

$$H_p = \text{Encoder}(p). \tag{2}$$

Next, we calculate the average similarity between post p and each symptom template:

$$Sim_p = \text{mean}(Sim(H_p, H_i)), \text{for } i = \text{A to I}, \tag{3}$$

where Sim_p represents the depression risk level of post p. We only use the LLM to annotate posts with the top $k\%$ of Sim_p scores, while the depression symptom vector E_p for all other posts is directly set to a zero vector. Here k is a hyperparameter. Finally, for each user u, we average all their depression symptom vectors to obtain their diagnostic criteria feature F_u^{DC}:

$$F_u^{DC} = \frac{1}{N}\sum_{p=1}^{N} E_p, \tag{4}$$

where N is the total number of posts by user u.

3.2.2 Mood Course Representation Construction

Mood course, defined as the temporal pattern and progression of emotional states [10], is critical in diagnosing clinical depression [13]. It delineates the onset, duration, and recurrence of mood episodes, providing insights into the disorder's nature and trajectory [17]. Accurately modeling mood course is pivotal for distinguishing depressive disorders from transient mood fluctuations, facilitating early detection and appropriate intervention [33].

Former works on depression detection has largely overlooked the mood course, focusing instead on static mood snapshots. Our study bridges this gap by explicitly modeling mood course and integrating it into our classification system. Next, we detail our approach.

Posts Filtering Not all posts are emotionally charged. We begin by filtering posts with a high emotional content. Following [28], we categorize emotions into five main types: 1) anger, 2) disgust, 3) anxiety, 4) happiness, and 5) sadness. For each of these emotional categories, we establish a template of emotional expressions. For instance, the template for sadness, T_5^E, is defined as:

"I am sad, sorrowful, melancholic, in pain, lost, depressed, pessimistic, tearful, grieving, mournful, depressed, suicidal, heartbroken, devastated, upset, crying, deeply saddened, disconsolate, dejected, lamenting, desolate, gloomy, mournful, weeping bitterly, desperate, heartbroken, indignant."

For each emotion template, we generate a representation using a pre-trained embedding model:

$$H_j^E = \text{Encoder}(T_i^E), \text{for } j = 1, 2, ..., 5. \tag{5}$$

Representation Construction For each post p, we obtain its embedding H_p. We calculate the similarity between post p and each emotion template as:

$$Sim_{pj} = Sim(H_p, H_j), \text{for } j = 1, 2, ..., 5. \tag{6}$$

For each emotion j, we retain posts within the top $m\%$ of similarity, forming the set S_j. Here m is a hyperparameter. The final set of posts with high emotional content, S, is the union of all S_j.

For each user u, we intersect their historical post set P_u with the high emotional content set S to obtain P_u^E. Based on this subset of emotionally expressive posts, we use an LLM to synthesize a description of user u's mood course, T^{MC}, using the following prompt:

"Assuming you are a psychiatrist specializing in depression, please summarize the mood course over this period based on the blogger's self-expressions:

Time: t_1, Post: p_1, Time: t_2, Post: p_2, ..."

We then compute the embedding of T^{MC}:

$$H^{MC} = \text{Encoder}(T^{MC}). \tag{7}$$

The mood course representation F_u^{MC} for user u, is calculated as:

$$F_u^{MC} = \frac{1}{|P_u^E| + 1} \left(\sum_{p \in P_u^E} H_p + H^{MC} \right), \tag{8}$$

where $\left|P_u^E\right|$ is the total number of posts in P_u^E.

3.2.3 Post History Representation Construction

In the sections above, we construct the diagnostic criteria feature and mood course representation from users' historical posts through filtering and labeling. While these filtering steps emphasize aspects crucial for medical diagnosis of depression, they may also result in information loss. To address this, we constructed a representation of the user's post history as follows:

$$F^{PH} = \frac{1}{N} \sum_{p=1}^{N} H_p, \tag{9}$$

where H_p is the embedding of the p-th post, and N is the total number of posts by the user.

3.2.4 Training and Predicting

In this section, we describe our training and predicting methodology. First, we integrate the various features. F^{MC} and F^{PH}, which shares the same space, are directly summed to avoid increasing dimension and exacerbating the risk of overfitting. Conversely, F^{DC} resides in a distinct space, and thus we concatenate this feature with the sum of the first two parts. The final feature vector F is obtained as follows:

$$F = \text{Concat}\left(F^{MC} + F^{PH}, F^{DC}\right) \tag{10}$$

We employ a Gradient Boosting Trees (GBT) approach for classification. This method involves constructing an ensemble of decision trees in a sequential manner, where each subsequent tree aims to correct the errors of its predecessor. It excels in automatically performing feature interactions by selecting optimal split criteria within its decision trees, thus effectively fusing the components of F. The final prediction, denoted as y, is a binary classification result derived from the ensemble model. The process is formalized as follows:

First, the ensemble model G is initialized and then enhanced iteratively by adding decision trees:

$$G_m(x) = G_{m-1}(x) + \nu \cdot h_m(x). \tag{11}$$

Each tree $h_m(x)$ is fitted to the negative gradient of the loss function evaluated at G_{m-1}, aiming to minimize:

$$\sum_{i=1}^{N} L\left(y_i, G_{m-1}\left(x_i\right) + h_m\left(x_i\right)\right). \tag{12}$$

The prediction y is given by the sign of $G_M(x)$, the output after M iterations:

$$y = \text{sign}\left(G_M(x)\right). \tag{13}$$

Here, L represents the loss function, N is the number of samples, M is the total number of trees, and ν is the learning rate.

3.2.5 Design for Explainability

Depression detection is a critical and safety-sensitive task, demanding both accuracy and explainability [45]. Existing work has revealed the potential of large models in interpretable mental health analysis [41, 43]. However, directly analyzing raw texts with LLMs can lead to accuracy discrepancies with domain-specific models and highly prompt-sensitive predictions [18]. In our approach, to ensure accuracy and stability, we do not rely on LLMs for direct judgment but instead utilize traditional classifiers. However, we also leverage the capabilities of LLMs to provide reasonable explanations for our model's decisions.

In our system, we employ LLMs to annotate symptoms of depression manifested in texts posted on social media platforms by users. Additionally, we generate descriptions of users' mood courses. These outputs serve not merely as intermediate variables in the system's operation but also as part of the final output, enhancing users' understanding of the system's results. Furthermore, we use the LLM to generate a new output, T^Exp, offering explanations for the model's classification results. The prompt used is as follows:

"Assuming you are a psychiatrist specializing in depression. Here is a user's mood course: T^{MC}; below are posts from this user displaying symptoms of depression and the types of symptoms exhibited: ...; this user has been determined by an automated depression detection system to be depressed/normal. Please consider the user's mood course and posts to generate an explanation for this judgment."

This explanatory output, T^{Exp}, along with the system's annotations of depressive symptoms in tweets, user mood course descriptions, and the system's classification results, constitute the final output of our system.

4 Experiments

4.1 Experimental Setup

4.1.1 Implementation Details

In our implementation, following [29, 7], we utilize cosine similarity to calculate the similarity between embeddings. To enhance the usability of our method in low-compute resource settings, we employ the low-resource-demanding pre-trained embedding model, gte-small-zh [5], which operates smoothly with just 1GB of memory. The embedding model in our system can easily be switched to other higher-performance models to further improve performance. For the LLM, we use GPT-3.5-Turbo-1103 [6], which requires only an internet connection to interact with it through the API service provided by OpenAI. The LLM in our system can also be substituted with open-source models, such as Mentallama [43], to further reduce monetary costs. Overall, our system can run at low computational costs, enhancing its accessibility. We utilize XGBoost [8] for an efficient implementation of Gradient Boosting Trees.

4.1.2 Dataset

We utilize a large-scale, open-source dataset, SWDD [7] to suppor our experiments. After expert screening, we keep 1000 depressed users and 19000 normal users, to simulate the actual depression rate. The dataset statistics are shown in Table 2. Training, validation, and test sets were divided in a 7:1:2 ratio, as recommended by [40]. We give a detailed discussion of the dataset in Section A.

[5]https://huggingface.co/thenlper/gte-small
[6]https://platform.openai.com/docs/models/gpt-3-5-turbo

	Depressed	Control
Num. of users	1000	19000
Num. of posts	69548	1314874
Avg. num. of posts per user	69.55	69.20

Table 2: Statistics of our utilized dataset.

Category	Method	Precision	Recall	F1-score	AUROC	AUPRC
Traditional Method	TF-IDF+XGBoost	0.3644	0.4300	0.3945	0.9023	0.4303
Deep Learning-Based Methods	HAN	0.5702	0.6500	0.6075	0.8929	0.5864
	Mood2Content	0.7216	0.7000	<u>0.7106</u>	<u>0.9537</u>	<u>0.7774</u>
PLM-Based Methods	FastText	<u>0.7467</u>	0.5600	0.6400	0.9441	0.6255
	gte-small	0.6359	0.6526	0.6200	0.9499	0.6959
	BERT	0.6667	0.6400	0.6531	0.9481	0.7102
	MentalRoBERTa	0.7326	0.6300	0.6774	0.9423	0.6880
LLM-Based Methods	ChatGPT	0.0875	0.7100	0.1559	0.6603	0.0767
	MentalLLama	0.0899	<u>0.7800</u>	0.1612	0.6821	0.0811
Our Method	DORIS	**0.7596**	**0.7900**	**0.7596**	**0.9715**	**0.8134**

Table 3: Performance of DORIS and baselines. The best scores are in bold, and second best scores are underlined.

4.1.3 Baseline Methods

We employed various baselines, including methods combining traditional feature extraction with classifiers: TF-IDF+XGBoost [31, 8, 40], deep learning approaches: HAN [44], PLM-based methods: FastText [21], BERT [15], MentalRoBERTa [20], and Mood2Content [7], as well as large language model methods like MentalLLama [43] and ChatGPT. For LLM-based approaches, we design prompts following [41]. Due to the token window constraint of large models, in some instances, we truncated the data to retain as many recent tweets as possible while ensuring the total prompt length did not exceed the limit.

4.1.4 Evaluation Metrics

Following prior works, we evaluate our method using five metrics: Precision, Recall, F1, AUROC, and AUPRC. **Precision** measures the proportion of true positive predictions in all positive predictions. **Recall** assesses the proportion of true positive predictions out of all actual positives. **F1 Score** is the harmonic mean of Precision and Recall, providing a balance between them. **AUROC** measures the probability that a classifier will rank a random positive instance higher than a random negative one. **AUPRC** is the area under the curve plotting precision against recall under various thresholds. It reflects the classifier's ability to identify positive instances among a set. For depression detection, AUPRC can be considered as the most important metric, as it best reflects the classifier's performance on highly imbalanced datasets [14].

4.2 Experimental Results

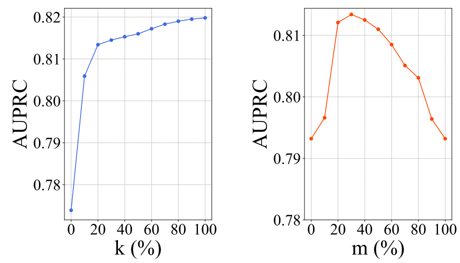
4.2.1 Overall Performance

In Table 3, we present a performance comparison between our method and comparison methods. The results highlight the superior performance of our approach. Specifically, we observe:

- Our method outperforms all baselines across all metrics. On AUPRC, which best reflects model performance on imbalanced datasets, DORIS shows an absolute improvement of 0.0360 over the best-performing baseline, Mood2Content. Additionally, in our implementation, we utilized gte-small and ChatGPT as the backbone. Compared to these models, our method significantly improves classification performance, demonstrating the effectiveness of our design.
- Models specifically designed for depression detection tasks outperform general models. Among various models, those designed for depression detection or mental health analysis, namely Mood2Content, MentalRoBERTa, and MentalLLama, achieved better performance. Our DORIS method, integrating medical understanding of depression and actual diagnostic practices, achieved the best results through task-specific design.

	F1-score	AUROC	AUPRC
Full Design	0.7596	0.9715	0.8134
w/o DC Feature	0.6867	0.9679	0.7739
w/o MC Representation	0.7415	0.9660	0.7932
w/o PH Representation	0.7200	0.9660	0.7817

Table 4: Experimental results of ablation study. DC denotes diagnostic criteria, MC denotes mood course, and PH denotes post history.



(a) The impact of k. (b) The impact of m.

Figure 3: Results of hyperparameter study.

- Direct classification with LLMs yields low performance. Our experiments show that both ChatGPT and MentalLLama underperformed, lagging significantly behind DORIS and other baselines on the crucial AUPRC metric. Methods based on LLMs exhibit high Recall but lower performance on other metrics, primarily due to their tendency to classify users as having depression. In our experiments, some truncation of input to LLMs was necessary, yet this truncation is not the cause of the poor performance of LLM-based methods, as it does not alter their inclination to predict the positive class. Our approach effectively harnesses the capabilities of LLMs without directly relying on their output for judgments, thereby achieving optimal results.

4.2.2 Ablation Study

In our method, the GBT classifier utilizes three constructed features to produce the final prediction results. Here we individually remove each of these feature sets from the classifier's input and observe the corresponding performance changes. By this means, we aim to validate the contribution of these features to the overall performance gain. The experimental results are presented in Table 4.

It is evident that removing any one of these three features results in a significant decrease in our system's performance. The most notable decline occurs upon removing the diagnostic criteria feature, with AUPRC dropping to 0.7739, which is lower than Mood2Content's 0.7774. This highlights the importance of the diagnostic criteria feature. Similarly, the performance significantly declines upon the removal of the other two features, proving the effectiveness of every part of DORIS's design.

4.2.3 Hyperparameter Study

In our method design, we introduce two hyperparameters: k, which controls the proportion of text retained with a high possibility of depression symptoms, and m, which controls the proportion of text retained with high emotional intensity. In this section, we explore the impact of these two parameters on model performance. In our experiments, k and m were varied from 0 to 100 in increments of 10, corresponding to retaining 0% to 100% of the texts. The experimental results are shown in Figure 3.

- **Impact of k.** The parameter k determines the proportion of text retained with a high risk of depression. As k increases, the AUPRC monotonically increases. This is because the more text is annotated, the more information is provided. Annotating 20% of the text with the LLM, compared to not annotating at all, significantly improves AUPRC, and it is already close to the performance when all texts are annotated. This demonstrates the rationality of our design for efficient implementation.
- **Impact of m.** The parameter m determines the proportion of text retained with high emotional intensity. As m increases, the AUPRC initially rises and then falls. The selection of m needs to

Diagnostic Criteria Feature Construction

Post Time	Text	Symptoms
April 23rd, 2015	The incident from two years ago, I'll never forget it for the rest of my life.	None
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.	Psychomotor agitation
May 3rd, 2015	I'm going back to school tomorrow, and I haven't touched volleyball in a month.	None
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.	Depressed mood
…	…	…
July 10th, 2015	I didn't realize I could lose so much weight without deliberately dieting.	Weight loss
July 16th, 2015	I really dislike people who lack integrity, and I find it very repulsive.	None
July 22th, 2015	I just want to find a place to hide and cry my heart out right now.	Depressed mood

Mood Course Representation Construction

Post Time	Text with High Emotional Content
April 30th, 2015	Driven to the brink of agitation, consumed by an overwhelming restlessness.
May 8th, 2015	The bitterness in my heart, no one understands. Even if I cry until tears run dry.
…	…
July 16th, 2015	I really dislike people who lack integrity, and I find it very repulsive.
July 22th, 2015	I just want to find a place to hide and cry my heart out right now.

Mood Course Description

The individual's social media expressions from April to July 2015 depict a concerning trajectory of emotional distress. Initially, they report intense agitation and restlessness, suggesting acute emotional turmoil. As weeks progress, there's a deepening sense of isolation and misunderstood bitterness, indicating potential depressive symptoms. Episodes of wanting to withdraw socially and expressions of disdain towards perceived moral failings in others hint at both a heightened sensitivity to interpersonal dynamics and a profound sense of disillusionment. The culmination of these feelings into a desire for seclusion and emotional release by mid-July underscores a significant emotional struggle, possibly pointing towards depressive episodes or acute stress reactions. This mood course, characterized by fluctuations in emotional intensity and themes of alienation, distress, and a quest for solace, warrants professional attention for potential mental health interventions.

Generated Explanation

Based on the user's mood course and specific posts, the judgment of depression appears to be accurate. The reported symptoms—psychomotor agitation, a persistently depressed mood, unintentional weight loss, and a desire for isolation—are hallmark indicators of depression. These symptoms not only cover a range of depressive disorder criteria, including mood disturbances, physical changes, and social withdrawal, but also span several months, suggesting a sustained period of emotional distress rather than a transient phase. The pattern and intensity of symptoms described align with clinical understanding of depressive episodes, warranting a diagnosis of depression rather than a normal fluctuation of emotions.

Figure 4: A case study of DORIS's output.

balance between sufficient filtering effect and not losing too much information. However, regardless of the value of m, constructing a mood course representation always benefits performance. This validates the rationality of our method design.

4.2.4 Case Study

In our system, we annotate symptoms exhibited in users' posts, summarized the users' mood course, and based on these, provide explanations for the system's final decision output. An example is presented in Figure 4 to aid in better understanding the operation of our system.

As seen, during the diagnostic criteria feature construction stage, our method accurately identifies depression-related symptoms in posts. In the mood course representation construction stage, our approach selects posts with high emotional density, thereby generating a mood course description. By combining identified symptoms of depression and the user's mood course, the system can generate a rational explanation for its judgment.

5 Conclusion and Future Work

In this work, we introduce a novel depression detection system named DORIS, which incorporates clinical practices in depression diagnosis, including the widely used diagnostic criteria DSM-5, and analyses of mood course, achieving high accuracy and interpretability in depression detection. Extensive experimental results demonstrate the effectiveness of our method, as well as the rationality of each design component in DORIS.

In the future, we plan to deploy our system in real-world settings to collaborate with human experts in clinical diagnosis. This will enable us to identify areas for improvement in our system and continue its iterative design.

Limitations

- **Our experiment is based on a single dataset.** Large-scale datasets with expert annotations are challenging to acquire in the field of depression detection, and we find only one dataset that meets our criteria, further refined based on our standards. The experimental results on this dataset have demonstrated the effectiveness of our method. However, we believe that experiments on more datasets could further validate our approach.
- **Our focus is solely on depression.** Mental health analysis and mental illness detection are crucial for human well-being. Our work concentrates on only one of these tasks: depression detection. While depression is a major mental illness, extending our method to detect other mental health illnesses, such as anxiety and bipolar disorder, would be more meaningful. Specifically, our explicit modeling of the mood course could be significantly beneficial for accurate and interpretable detection of bipolar disorder. In the future, we plan to extend our method to a broader range of mental health analysis and mental illness detection tasks.

Ethical Considerations

The texts presented in our paper include content with depressive tendencies, which may cause discomfort. The dataset used is a public dataset, and we have strictly adhered to the guidelines provided by the dataset providers. To protect user privacy and prevent misuse, all examples shown in the paper have been moderately obfuscated [5]. In using the API services provided by OpenAI, we have complied with OpenAI's terms and policies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Mihael Arcan, Paul-David Niland, and Fionn Delahunty. An assessment on comprehending mental health through large language models. *arXiv preprint arXiv:2401.04592*, 2024.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[4] Shani Bril-Barniv, Galia S Moran, Adi Naaman, David Roe, and Orit Karnieli-Miller. A qualitative study examining experiences and dilemmas in concealment and disclosure of people living with serious mental illness. *Qualitative Health Research*, 27(4):573–583, 2017.

[5] Amy Bruckman. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4:217–231, 2002.

[6] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, and Victor Carneiro. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research*, 21(6):e12554, 2019.

[7] Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538, 2023.

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[9] Raymond Chiong, Gregorious Satia Budhi, and Sandeep Dhakal. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems*, 36(6):99–105, 2021.

[10] AL Cochran, MG McInnis, and DB Forger. Data-driven classification of bipolar i disorder from longitudinal course of mood. *Translational psychiatry*, 6(10):e912–e912, 2016.

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

[12] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

[13] E Jane Costello, Daniel S Pine, Constance Hammen, John S March, Paul M Plotsky, Myrna M Weissman, Joseph Biederman, H Hill Goldsmith, Joan Kaufman, Peter M Lewinsohn, et al. Development and natural history of mood disorders. *Biological psychiatry*, 52(6):529–542, 2002.

[14] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

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

[16] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[17] Álvaro Frías, Daniel P Dickstein, John Merranko, Mary Kay Gill, Tina R Goldstein, Benjamin I Goldstein, Heather Hower, Shirley Yen, Danella M Hafeman, Fangzi Liao, et al. Longitudinal cognitive trajectories and associated clinical variables in youth with bipolar disorder. *Bipolar disorders*, 19(4):273–284, 2017.

[18] Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*, 2024.

[19] Julia Ive, George Gkotsis, Rina Dutta, Robert Stewart, and Sumithra Velupillai. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 69–77, 2018.

[20] Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*, 2023.

[21] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[22] Murad Moosa Khan et al. Economic burden of mental illnesses in pakistan. *Journal of Mental Health Policy and Economics*, 19(3):155, 2016.

[23] E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. # mydepressionlookslike: Examining public discourse about depression on twitter. *JMIR mental health*, 4(4):e8141, 2017.

[24] Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 407–411, 2020.

[25] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages*, pages 28–39. Springer, 2016.

[26] Anshu Malhotra and Rajni Jindal. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, page 109713, 2022.

[27] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572, 2022.

[28] Keith Oatley and Philip N Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50, 1987.

[29] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 88–97, 2018.

[30] David Owen, Jose Camacho Collados, and Luis Espinosa-Anke. Towards preemptive detection of depression and anxiety in twitter. *arXiv preprint arXiv:2011.05249*, 2020.

[31] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[32] Darrel A Regier, Emily A Kuhl, and David J Kupfer. The dsm-5: Classification and criteria changes. *World psychiatry*, 12(2):92–98, 2013.

[33] Karen D Rudolph, Constance Hammen, and Shannon E Daley. Mood disorders. *Behavioral and emotional disorders in adolescents: Nature, assessment, and treatment*, pages 300–342, 2006.

[34] Nicolas Rüsch, Matthias C Angermeyer, and Patrick W Corrigan. Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European psychiatry*, 20(8):529–539, 2005.

[35] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. 2017.

[36] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[37] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893, 2019.

[38] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[40] Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. Exploring social media for early detection of depression in covid-19 patients. *arXiv preprint arXiv:2302.12044*, 2023.

[41] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, 2023.

[42] Kailai Yang, Tianlin Zhang, and Sophia Ananiadou. A mental state knowledge–aware and contrastive network for early stress and depression detection on social media. *Information Processing & Management*, 59(4):102961, 2022.

[43] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*, 2023.

[44] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[45] Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46, 2022.

[46] Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. Phq-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, 60(5):103417, 2023.

[47] Tianlin Zhang, Kailai Yang, and Sophia Ananiadou. Sentiment-guided transformer with severity-aware contrastive learning for depression detection on social media. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 114–126, 2023.

[48] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

[49] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*, 2021.

A Discussions on Our Utilized Dataset

Unlike other machine learning tasks with large-scale, high-quality datasets with golden labels, acquiring well-labeled, large-scale datasets for depression detection is challenging due to the sensitivity and privacy concerns associated with mental illnesses. Some datasets utilize self-reported medical diagnoses as the criterion for identifying users with depression. The advantage of this approach is its relatively high reliability and ease of data collection without needing expert involvement. However, this leads to datasets comprising users who are already medically diagnosed with depression and willing to discuss it online. However, the most valuable aspect of depression detection task is assisting human experts in identifying potentially undiagnosed users on the internet and intervening when appropriate. Moreover, posts from self-disclosing users often contain explicit mentions of depression or medication discussions, making identification relatively easy. We want our utilized dataset to include users who have not self-reported depression but are highly likely to be depressed, demonstrating the effectiveness of our method in such scenarios. Additionally, the ratio of depressed to normal users in most previous datasets [25, 49, 40] is around 1:4 or even 1:1, deviating far from the actual prevalence of about 5% in the population. To simulate a real-world application scenario, our dataset has a ratio of 1:19 depressed to normal users. Experiments on such a dataset can more adequately validate the practical performance of various methods.

Given the scarcity of large-scale open-source datasets in depression detection, finding a dataset that meets our criteria was challenging. Fortunately, we discovered a large-scale, open-source dataset, SWDD [7], supporting our experiments. This high-quality dataset has been carefully annotated by experts, including users who self-disclosed a medical diagnosis of depression and those identified by experts as highly likely to be depressed based on medical criteria, and without self-disclosure. We retain these non-self-disclosing users and, after further expert screening, keep 1000 depressed users and 19000 normal users. Following the findings of [16], that utilizing posts from the six months preceding a medical diagnosis, rather than the entire history, yields a higher accuracy in predicting clinically diagnosed depression; and due to the token window constraint of LLMs, we retain only the posts posted within six months prior to each user's last post. The dataset statistics are shown in Table 2. Training, validation, and test sets were divided in a 7:1:2 ratio, as recommended by [40]. The original language of this dataset is Chinese. In our paper, to make our work accessible to a broader audience, all texts from the dataset are translated into English when presented.

B Symptom Templates

A. Depressed mood. I feel low, unhappy, joyless, depressed, oppressed, gloomy, disappointed, melancholic, sad, distressed, heartbroken, a sense of loss, often feeling heavy-hearted, experiencing despair and despondency, always feeling sorrowful with an urge to cry, experiencing inner pain and emptiness.

B. Loss of interest/pleasure. I have lost interest, feel indifferent, bored, unconcerned, lack enthusiasm, am unmotivated, have no interest in activities, am unmotivated, find almost everything uninteresting, lack motivation, find significantly reduced pleasure, cannot experience happiness, feel the world is dull, and cannot muster energy all day.

C. Weight loss or gain. I experience reduced appetite, often feel full, lack of appetite, nausea, abnormal weight loss, difficulty swallowing, emaciation, loss of appetite, poor appetite, weight loss, or abnormal weight gain, sudden weight increase, unexplained weight gain.

D. Insomnia or hypersomnia. I suffer from sleep disorders, depend on sleeping pills, often experience insomnia, have difficulty falling asleep, rely on sleep medication, frequently stay up late, struggle with sleep difficulties, and exhibit symptoms of insomnia, tossing and turning at night, or hypersomnia, oversleeping, sleep excess, prolonged sleep duration, or excessive sleepiness.

E. Psychomotor agitation or retardation. I am neurotic, easily agitated, emotionally unstable, impatient, anxious, restless, mentally tense, irritable, often feeling mentally uneasy and agitated, fidgety, displaying impulsive and irritable behavior, and my emotions are easily out of control.

F. Fatigue I feel fatigued, listless, exhausted, physically weakened, lacking in energy, dispirited, frequently tired, powerless, often feeling weary, unable to muster strength, feeling a heavy body, lacking in vitality and vigor, always feeling drowsy and lethargic.

G. Inappropriate guilt. I have feelings of self-denial, lack of confidence, self-doubt, inferiority, disappointment, guilt, negative self-evaluation, self-blame, frequently belittle myself, feel incompetent and worthless, believe that I have achieved nothing and am a failure, feel disappointed in my expectations of myself and my family, often feel guilty and blame myself, thinking that everything is my fault.

H. Decreased concentration. I experience slow thinking, difficulty concentrating, reduced judgment, memory decline, distractibility, indecision, scattered attention, difficulty thinking, lack of focus, difficulty paying attention, decreased cognitive ability, hesitancy in making decisions, often feeling mentally spaced out, unable to concentrate.

I. Thoughts of suicide. I have a desire for death, self-harming behavior, suicidal thoughts, thoughts of ending my life, suicidal actions, thoughts of suicide, self-injury, recurring thoughts of death, suicidal tendencies, self-mutilation, cutting wrists with blades, jumping from heights to commit suicide, overdosing to commit suicide, making plans for suicide.

C Emotion Templates

1) Anger I am angry, mad, agitated, annoyed, indignant, irritable, furious, disgusted, incensed, enraged, irritated, vexed, resentful, in a rage, glaring, shouting, screaming, insulting, hating, bellowing, outraged, ranting, detesting, fuming, and uncontrollably angry.

2) Disgust I detest, loathe, disgust, abhor, hate, tire of, feel nauseated by, have a strong aversion to, despise, scorn, disdain, reject, find repugnant, utterly dislike, disdain, feel revulsion, despise, dislike intensely, abominate, have a strong displeasure, grow weary of, become impatient with, dismiss, look down upon, and utterly abhor.

3) Anxiety I feel anxious, uneasy, worried, concerned, nervous, restless, panicked, fretful, afraid, uncertain, apprehensive, tense, jittery, indecisive, fearful, flustered, melancholic, frightened, apprehensive, full of doubts, brooding, terrified, distrustful, terrified, and on edge.

4) Happiness I am happy, joyful, glad, blissful, merry, satisfied, delighted, elated, pleased, laughing, cheerful, excited, jubilant, optimistic, enthusiastic, cheerful, uplifted, exuberant, overjoyed, jubilant, with a smile on my face, pleasantly surprised, beaming with joy, and my heart blooms with happiness.

5) Sadness I am sad, sorrowful, melancholic, in pain, lost, depressed, pessimistic, tearful, grieving, mournful, depressed, suicidal, heartbroken, devastated, upset, crying, deeply saddened, disconsolate, dejected, lamenting, desolate, gloomy, mournful, weeping bitterly, desperate, heartbroken, indignant.