Detecting Bipolar Disorder from Misdiagnosed Major Depressive Disorder with Mood-Aware Multi-Task Learning

Daeun Lee[1*], **Hyolim Jeon**[1*], **Sejung Son**[1], **Chaewon Park**[1],
Jihyun An[2], **Seungbae Kim**[3], **Jinyoung Han**[1†]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, South Korea
[2] Department of Psychiatry, Samsung Medical Center, Seoul, South Korea
[3]Computer Science & Engineering Department, University of South Florida, Tampa, FL, USA
{delee12, jinyounghan}@skku.edu, jh85.an@samsung.com
{gyfla1512, maze0717, chaewonpark}@g.skku.edu, seungbae@usf.edu

Abstract

Bipolar Disorder (BD) is a mental disorder characterized by intense mood swings, from depression to manic states. Individuals with BD are at a higher risk of suicide, but BD is often misdiagnosed as Major Depressive Disorder (MDD) due to shared symptoms, resulting in delays in appropriate treatment and increased suicide risk. While early intervention based on social media data has been explored to uncover latent BD risk, little attention has been paid to detecting BD from those misdiagnosed as MDD. Therefore, this study presents a novel approach for identifying BD risk in individuals initially misdiagnosed with MDD. A unique dataset, *BD-Risk*, is introduced, incorporating mental disorder types and BD mood levels verified by two clinical experts. The proposed multi-task learning for predicting BD risk and BD mood level outperforms the state-of-the-art baselines. Also, the proposed dynamic mood-aware attention can provide insights into the impact of BD mood on future risk, potentially aiding interventions for at-risk individuals.

1 Introduction

Bipolar Disorder (BD) is a mental disorder characterized by recurring intense mood swings, ranging from depression to manic. Unfortunately, individuals with BD are at a higher risk of suicide than the general population (Pompili et al., 2013) and those with other psychiatric disorders (Rihmer and Kiss, 2002). Hence, accurate and early detection of individuals suffering from BD is vital to ensure timely suicide prevention.

According to previous studies, 17% to 50% of BD is misdiagnosed as *Major Depressive Disorder (MDD)* (Passos et al., 2019; Angst et al., 2011; Daveney et al., 2019), characterized by persistent sadness and a loss of interest (American Psychiatric Association et al., 2013). While depressive episodes in BD are similar to those in MDD, critical distinction arises from manic episodes, which include an elevated mood, impulsivity, and racing thoughts (de Almeida and Phillips, 2013). For instance, individuals with MDD typically exhibit stable low mood variability, whereas those with BD have unstable mood swings due to recurrent mood episodes (Ortiz et al., 2015). Nevertheless, individuals with BD often tend not to perceive manic episodes as abnormal, and some could even find these states preferable (Baldessarini et al., 2010), leading them to recognize themselves as MDD. Also, more than 58% of BD patients undergo a depressive episode before experiencing a manic episode and predominantly seek help during the depressive phase (Etain et al., 2012). Since BD and MDD require distinct treatments and prognoses, misdiagnosis can lead to significant delays in receiving proper treatment and can also increase the risk of suicide (Keramatian et al., 2022). Therefore, precise identification of patients who may develop BD after an initial diagnosis of MDD is of great clinical importance (Hu et al., 2020).

Importantly, due to a huge gap in understanding prognosis trajectories between real-world scenarios and clinical settings, where clinicians' interaction with patients is limited and relies on patients' subjective accounts (Harvey et al., 2022), there is an emerging interest in utilizing real-world data from non-clinical contexts like social media to gain insights into distinct behavioral traits that can be linked to BD and MDD (Jagfeld et al., 2021; Hwang and Hollingshead, 2016). However, while several studies have suggested methods for analyzing and categorizing mental states, including BD and MDD, using social media data (Kim et al., 2020; Cohan et al., 2018; Gkotsis et al., 2016), little attention has been paid to detecting a shift in diagnosis from MDD to BD that can be crucial in identifying individuals who have developed BD after an initial diagnosis of MDD. Instead of a one-

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4954–4970
June 16-21, 2024 ©2024 Association for Computational Linguistics

time diagnosis in a clinical setting, a behavior trait of an individual who develops BD after a diagnosis of MDD can be comprehensively captured and analyzed using longitudinal social media data that includes dynamic mood variation across time.

Therefore, this study aims to detect BD risk among patients incorrectly diagnosed with MDD or wrongly perceived as MDD, using their historical mood fluctuations revealed in their past social media activity. To this end, we introduce a novel dataset, *BD-Risk*, that includes (i) the types of mental disorders (e.g., BD or MDD) and (ii) BD mood level on a scale from -3 to 3, which are verified by clinical experts. Our data analysis indicates a strong correlation between mood variation and the BD risk of users, which is in line with the DSM-5 (American Psychiatric Association et al., 2013), which emphasizes the importance of evaluating not only the absolute mood level but also the relative mood variation to monitor the risk of BD. Thus, a multi-task learning model is proposed in which three tasks are simultaneously learned: (i) predicting the BD risk for users initially diagnosed with MDD, (ii) estimating BD mood level, and (iii) assessing mood variation from their posts over time.

Our extensive experiments demonstrate that the proposed model outperforms all the baseline models in BD risk prediction. We find that jointly learning BD mood information can improve performance by transferring informative representations and parameters among the tasks. Additionally, the attention scores provided by the proposed dynamic mood-aware attention method can provide valuable insights into comprehending the influence of BD mood on a user's future BD risk. This capability has the potential to assist clinicians in deepening their comprehension of the connections between psychiatric conditions, thereby enabling appropriate interventions, including the management of individuals undergoing depressive episodes (Ratheesh et al., 2017). We summarize the contributions of this work as follows.

- We release our codes and a novel BD dataset, *BD-Risk*[1], which contains both mental disorder diagnosis and BD mood levels, validated by clinical experts. The dataset can benefit researchers in developing mental disorder prevention.
- To the best of our knowledge, this is the first study that proposes a multi-task learning model for (i) predicting the BD risk for users initially diagnosed with MDD, (ii) estimating BD mood level, and (iii) assessing mood variation. The model can accurately capture BD mood level and variation, outperforming the state-of-the-art methods in detecting BD risk at an early stage.
- The proposed dynamic mood-aware attention method provides interpretability that can help clinicians track patients' mood swings, thereby giving early interventions by identifying their likelihood to transition from MDD to BD.

[1]Data and Code are available at: https://github.com/DSAIL-SKKU/Detecting-BD-from-Misdiagnosed-MDD_NAACL_2024

2 Related Work

2.1 Social Media Datasets for MDD and BD

As social media has become a vital platform for individuals sharing their daily experiences and emotions (Lee et al., 2023), there is an increasing interest in creating datasets using social media on diverse mental disorders (Cohan et al., 2018; Coppersmith et al., 2014) such as MDD (Shao et al., 2019; Losada et al., 2019) and BD (Lee et al., 2023; Sekulić et al., 2018; Jagfeld et al., 2021). However, while prior datasets can be useful in mental health research, little attention has been paid to developing a dataset that contains information about diagnosis shifts from MDD to BD, which is crucial for providing valuable insights into early intervention for patients who could develop BD after an initial diagnosis of MDD (Hu et al., 2020). Table 1 compares the existing popular BD datasets and the proposed BD-Risk. One of the potential limitations of existing MDD and BD datasets is the reliability due to insufficient validation and verification. As shown in Table 1, existing datasets categorize users based on subreddit topics where they posted (Kim et al., 2020) or using computational keyword matching (e.g., from *"I am diagnosed with bipolar"*) (Guo et al., 2021; Jagfeld et al., 2021; Cohan et al., 2018; Sekulić et al., 2018), which may not be accurate. Many users who initially reported being diagnosed with MDD disclosed later a BD diagnosis on social media, possibly due to medical misdiagnosis or their misperception (Jagfeld et al., 2021). This paper proposes a novel dataset, *BD-Risk*, that contains information about diagnosis shifts from MDD to BD. As shown in Table 1, our dataset stands alone in encompassing not only diagnosis for BD and MDD but also detailed BD mood information, validated by psychiatrists. We make this valuable dataset publicly accessible, firmly believing it will

Table 1: Comparisons between the proposed BD-Risk and other BD datasets.

Dataset	BD-Risk (Ours)	Jagfeld et al. (2021)	Sekulić et al. (2018)
BD diag.	✓	✓	✓
MDD diag.	✓	✓	✗
BD Mood	✓	✗	✗
Publicly Available	✓	✓	✗
Expert Validation	✓	✗	✗
Social Media	Reddit	Reddit	Reddit
# of users	1,025	19,685	3,488
# of posts	7,346	21,407,595	-

be a substantial contribution to both the mental health and machine learning communities.

2.2 Detecting BD or MDD on Social Media

Many researchers have utilized social media data to explore the characteristics of BD (e.g., linguistic patterns (Sekulić et al., 2018), BD symptoms (Lee et al., 2023), emotion (Guo et al., 2021), and BD status (Sekulić et al., 2018)) and MDD (e.g., linguistic variance (Naseem et al., 2022a), non-verbal features (Yoon et al., 2022), and MDD-related features (Sadeque et al., 2018)). However, these studies examine each condition in isolation, missing the opportunity to model shared influential factors (e.g., depressive symptoms in both BD and MDD). Thus, prior researchers have demonstrated the benefits of multi-task learning, as it implicitly captures interactions among related mental health conditions (Lee et al., 2023; Azim et al., 2022; Lokala et al., 2022). Therefore, this paper proposes a multi-task learning model, concurrently addressing the following three tasks: (i) predicting BD risk in individuals misdiagnosed with MDD, (ii) estimating BD mood level, and (iii) assessing mood variation.

3 BD-Risk Dataset

Identifying individuals who may develop BD after being initially diagnosed with MDD is essential due to the potential risks of misdiagnosis, such as delayed treatment and increased suicide risk (Keramatian et al., 2022). Therefore, we aim to detect early signs of BD among people misdiagnosed with MDD. To this end, we build a dataset, *BD-Risk*, that includes the following labels: (i) specific mental disorders (e.g., BD or MDD) and (ii) the BD mood level scaling from -3 to 3. In this section, we provide a data collection approach, annotation strategy, and statistics of annotation evaluation.

Table 2: Summary of annotated labels in our data.

Type	Total	Category	Count (%)
Diagnosis	1,025 users	*MDD*	569 (55.5%)
		MDD → BD	456 (44.5%)
BD Mood Level	7,346 posts	3	28 (0.38 %)
		2	93 (1.27 %)
		1	338 (4.60 %)
		0	351 (4.78 %)
		-1	2,236 (30.44 %)
		-2	2,376 (32.34 %)
		-3	1,924 (26.19 %)

3.1 Data Collection

We collected posts from various subreddits related to MDD and BD, such as *r/Depression*, *r/bipolar*, *r/BipolarReddit*, and *r/BipolarSOs*, using the *Reddit* API[2] (Baumgartner et al., 2020), with the period from January 1st, 2008, to March 4th, 2023. We selected users who exclusively posted their first 3 or more posts on *r/Depression* before writing on BD-related subreddits. In total, our dataset consists of 7,346 posts written by 1,025 users as shown in Table 1.

3.2 Annotation Process

To label the dataset, we recruited three researchers as annotators who were knowledgeable in mental health and proficient in English. Under the guidance of a psychiatrist, trained annotators labeled anonymized posts via the open-source annotation tool *Doccano* (Nakayama et al., 2018). Note that we removed all personal identifiers from collected posts before assigning annotation tasks for ethical concerns. When there were conflicts among annotators, all annotators engaged in discussions and concluded by following the guidance of a psychiatrist. Annotation results are described in Table 2.

A. Diagnosis Types: According to DSM-5 (American Psychiatric Association et al., 2013), diagnosis types include MDD and BD with subtypes such as 'type-I', 'type-II', and 'Not Otherwise Specified BD'. To construct the dataset, we first searched phrases relevant to clinical diagnoses within the text (Jagfeld et al., 2021), such as *"I've recently been diagnosed with MDD officially"*. Subsequently, the annotators meticulously assigned appropriate diagnosis types to users who explicitly mentioned their treatment process with mental health experts. To ensure the inclusion of users with clinical diagnoses while annotating BD diagnoses, we intentionally selected posts containing

[2] https://www.reddit.com/dev/api/

keywords related to clinical evidence. For example, mentions of medication or hospital visits (i.e., objective clinical evidence) are included (e.g., *"I was diagnosed as bipolar at the hospital and prescribed medication"*), while excluding those lacking such indicators hence can be subjective (e.g., *"I am bipolar"*). Posts related to the diagnoses of individuals other than the users, such as family members or friends, were excluded from the dataset.

B. BD Mood Level: We employ the DSM-5 to assess the severity of mood (such as Mild, Moderate, or Severe) (American Psychiatric Association et al., 2013). Using this criterion, we assign post annotations with a scale ranging from -3 to 3. Negative values represent depressive moods, while positive values represent manic moods. We assume the posts exhibiting both manic and depressive moods are regarded as manic moods (American Psychiatric Association et al., 2013). For a comprehensive understanding of each category and example, please refer to Table 7 in the Appendix A.

C. Data Filtering: To alleviate the time-delay issue, where the reported diagnoses may not align with the timing of the posts, we implemented three data filtering strategies as follows. First, we excluded users who mentioned BD-related words (e.g., *"bipolar"*, *"bd"*) or posted in the BD-related subreddit, such as *r/BipolarReddit*, before reporting the BD diagnosis to determine the diagnosis transition's timing accurately. Second, while annotating BD diagnoses, we manually examined users' post content, retaining only those indicating recent or within a maximum of one year of diagnoses. We removed posts that did not specify the timing. Finally, we gathered posts where users reported a BD diagnosis in at least three posts within a specific time frame after reporting an MDD diagnosis. This is illustrated in Figure 3b, showing an average time difference of approximately 560 days between MDD and BD diagnosis.

3.3 Expert Validation

To ensure the accuracy of annotated labels, we conducted a validation process with domain experts, specifically a psychiatrist (E1) and a clinical psychologist (E2). We randomly selected 150 posts from 120 users for this evaluation. The reliability of the annotations was measured using Krippendorff's alpha-reliability (Krippendorff, 2018) and Cohen's Inter-Annotator Agreement (Cohen, 1960). Table 3 indicates a high level of agreement between experts and annotators, with an overall Krippendorff's α score of 0.87 and Cohen's κ score of 0.65, confirming the reliability of our dataset.

Table 3: Expert validation result of BD mood levels in *BD-Risk* (E1, E2: Clinical Experts /I: Annotators).

Cohen's κ	**E1**	**E2**	**I**
E1	1	-	-
E2	0.69	1	-
I	0.63	0.65	1
Krippendorff's α	0.87		

3.4 Class Generation

According to the annotation results, users can be classified into two groups $y_r \in \{MDD, MDD \rightarrow BD\}$: (i) those who were diagnosed as MDD only (*MDD*, 569 users), and (ii) those who were initially diagnosed or self-reported as MDD and later re-diagnosed as BD (*MDD* $\rightarrow$ *BD*, 456 users) (refer to Table 2). In Figure 1, for example, let $u_i \in U = \{u_1, u_2, \ldots, u_i\}$ represents a user who shared n posts $P_i = \{p_1^i, p_2^i, \ldots, p_t^i\}$ on social media, where t denotes the posting time. Both users, u_1, and u_3, were initially diagnosed with MDD. However, only in the case of u_3 the diagnosis later changed to BD, indicating an initial misdiagnosis. Therefore, u_1 is assigned to the *MDD*, while u_3 is classified as *MDD* $\rightarrow$ *BD*.

To examine the significance of assessing both the absolute mood level and the relative mood variation (American Psychiatric Association et al., 2013), a BD mood level y_m within the range of [-3, 3] as well as a mood variation level y_v within the range of [-6, 6] is assigned to each post p_t^i. y_v is calculated as the difference in mood levels between a user's n^{th} post and the 1^{st} post. For instance, if y_m for each post p_1^3 and p_2^3 are -3 and -2 for u_1, the corresponding y_v values would be 0 and +1, respectively, as shown in Figure 1.

3.5 Data Analysis

We analyze our dataset, *BD-Risk*, to understand the similarities and differences between *MDD* and *MDD* $\rightarrow$ *BD* groups. The analysis results are included in the Appendix A.1. In summary, our analysis reveals distinct linguistic patterns and mood expressions between the two groups in line with clinical trials. These findings offer valuable insights into the characteristics of individuals whose diagnoses shift from MDD to BD compared to those with MDD alone.

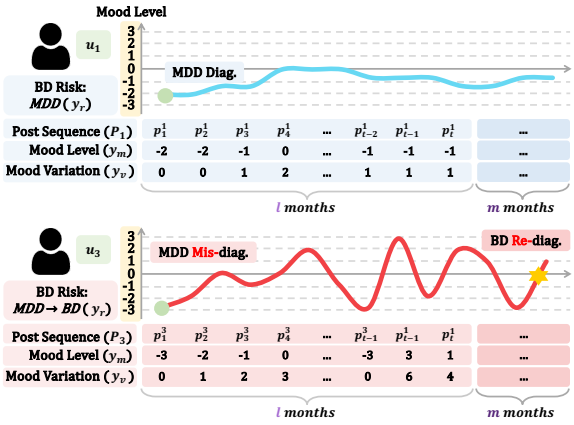


Figure 1: Example post sequences of Reddit users. u_1 belongs to the *MDD*, while u_3 belongs to the *MDD* → *BD* due to u_3's diagnosis shift from MDD to BD.

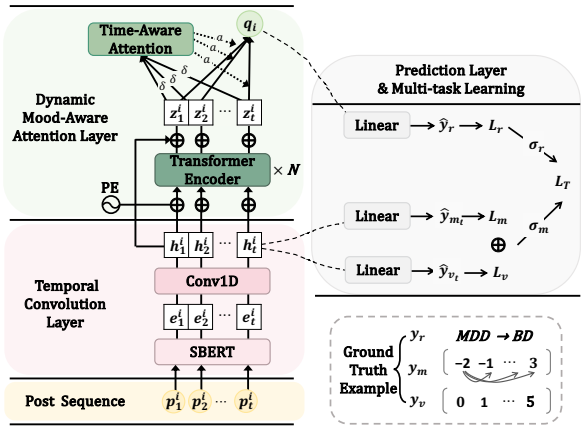


Figure 2: The overall architecture of the proposed multi-task learning model.

4 The Model

This section describes the problem definition and our proposed multi-task learning model that can (i) detect the BD risk for users initially diagnosed with MDD, (ii) estimate BD mood level, and (iii) assess mood variation from their posts over time. The model includes three main components: a Temporal Convolution layer, a Dynamic Mood-Aware Attention layer, and a Prediction layer. The overall architecture is illustrated in Figure 2.

4.1 Problem Definition

BD Risk Detection: Our main task is to identify the BD risk $y_r \in \{MDD, MDD \rightarrow BD\}$ for user u_i by analyzing the sequence of posts P_i as illustrated in Figure 1. Since BD patients often exhibit recurrent mood swings, a substantial duration is required to observe their mood patterns (Egeland et al., 2012). Therefore, we define a timeline that contains (i) the past l months for observation and (ii) the future m months for identifying BD diagnosis within a series of posts of a user. In other words, the time interval between p_1^i and p_t^i is l months. Further details on the performance of different post-sequence durations can be found in Figure 3a.

BD Mood Level & Mood Variation Estimation: The two auxiliary tasks involve estimating the BD mood level y_m ranging from [-3, 3] and evaluating the mood variation y_v ranging from [-6, 6] for each post p_t^i. For further elaboration on the label construction process, refer to § 3.4 *Class Generation*.

4.2 Temporal Convolution Layer

We extract a semantic representation for each post p_t^i using a pretrained Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which shows promising results in detecting future suicidality (Lee et al., 2023) and identifying mood change (Azim et al., 2022). SBERT incorporates siamese and triplet networks to generate sentence embeddings by calculating the average of output vectors for all tokens. We encode p_t^i as follows:

$$e_t^i = SBERT(p_t^i) \in \mathbb{R}^{n \times d} \quad (1)$$

where d is the dimension of text representations. Then, a 1-D convolution is used with a kernel size k to capture the dynamic mood patterns over time. This method inspects the neighboring elements in the sequential data and extracts local patterns of each sequence P_i (Ma et al., 2020).

$$h_t^i = Conv1D(e_t^i, k) \in \mathbb{R}^{n \times d} \quad (2)$$

4.3 Dynamic Mood-Aware Attention Layer

To obtain the historical context, the Transformer encoder (Vaswani et al., 2017) is applied with positional encoding (PE), which is widely recognized for its effectiveness in capturing global dependencies as follows:

$$z_t^i = h_t^i + Trans.(h_t^i + PE) \in \mathbb{R}^{n \times d} \quad (3)$$

$$PE = \begin{cases} PE[pos, 2j] = sin(pos/10000^{2j/d}) \\ PE[pos, 2j+1] = cos(pos/10000^{2j/d}) \end{cases} \quad (4)$$

where the Transformer encoder consists of multi-head attention and feed-forward layers. PE utilizes sine and cosine functions with frequencies determined by feature index and $pos = 1, \ldots, t$ and $j = 0, \lfloor \frac{2}{d} \rfloor$. While PE helps retain ordering information for sub-series, temporal information can

be inevitably lost due to the permutation invariant self-attention mechanism (Zeng et al., 2023). To address this issue and preserve temporal information in a sequence, we apply the time-aware parameter $\delta(z_t^i, \Delta_t)$ (Lee et al., 2023), fused with a self-attention mechanism to emphasize critical mood states that significantly influence the BD risk classification decision over time as follows:

$$q_i = \sum_{t=1}^{n} a_t^i z_t^i \quad (5)$$

$$a_t^i = \frac{exp(c^\top tanh(W \cdot \delta(z_t^i, \Delta_t) + b))}{\sum exp(c^\top tanh(W \cdot \delta(z_t^i, \Delta_t) + b))} \quad (6)$$

$$\delta(z_t^i, \Delta_t) = z_t^i + sigmoid(\theta - \mu\Delta_t) z_t^i \quad (7)$$

where $W \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, θ and μ are learnable parameters, and $tanh()$ is the activation function. θ and μ are influenced by z_t^i, and Δ_t is the time interval between the target post p_t^i and first post p_1^i. Given that mood can persist for hours or days (Davidson et al., 2009), $\delta(z_t^i, \Delta_t)$ can be useful in capturing the present state by evaluating the duration of past moods.

4.4 Prediction Layer

BD Risk Detection Decoder: To identify the BD risk for each sequence P_i, the proposed decoder generates the final prediction vector as follows:

$$\hat{y_{r_i}} = \mathcal{F}(tanh(\mathcal{F}(q_i))) \in \mathbb{R}^{1 \times |y_r|} \quad (8)$$

where $\mathcal{F}$ is a fully-connected layer.

BD Mood Level & Variation Estimation Decoder: By using temporal sequence embeddings h_t^i, the logits of labels for the BD mood level and mood variation can be inferred as:

$$\hat{y_{m_t^i}} = \mathcal{F}(tanh(\mathcal{F}(h_t^i)) \in \mathbb{R}^{1 \times |y_m|} \quad (9)$$

$$\hat{y_{v_t^i}} = \mathcal{F}(tanh(\mathcal{F}(h_t^i)) \in \mathbb{R}^{1 \times |y_v|} \quad (10)$$

4.5 Multi-task Learning

We train the model by learning a main task and two auxiliary tasks jointly. However, since each task has different scales (i.e., post-level mood prediction vs. sequence-level BD risk prediction), we apply the uncertainty weight loss (Kendall et al., 2018), which assigns weights to multiple loss functions based on task-dependent uncertainty. This method enables the model to learn effectively across various scales and units for different tasks simultaneously. The final loss L_T is derived as:

$$\mathcal{L}_T = \frac{1}{2\sigma_r^2}\mathcal{L}_r(W) + \frac{1}{2\sigma_m^2}(\mathcal{L}_m + \mathcal{L}_v)(W) + log\sigma_r\sigma_m \quad (11)$$

$$\mathcal{L}_r = -\sum_{i=1}^{b} y_{r_i} \log \hat{y_{r_i}} \quad (12)$$

$$\mathcal{L}_m = \sum_{i=1}^{b}\sum_{t=1}^{n} \left(\hat{y_{m_t^i}} - y_{m_t^i}\right)^2 \quad (13)$$

$$\mathcal{L}_v = \sum_{i=1}^{b}\sum_{t=1}^{n} \left(\hat{y_{v_t^i}} - y_{v_t^i}\right)^2 \quad (14)$$

where the cross-entropy loss is calculated for the main task, and the MSE loss is computed for each auxiliary task. Since each auxiliary task has the same granularity, the losses are summed up as an auxiliary task loss. σ_r and σ_m are the learnable parameters representing uncertainty for each task, W is the weight parameter, and b is the batch size.

5 Experiments

5.1 Experimental Settings

We ensure that users in the test set are entirely disjoint and do not overlap with those in the training set. For reproducibility, detailed experimental settings are summarized in Appendix B.

5.2 Baseline Models

We evaluate various baselines for comprehensive performance comparisons. As studies on identifying BD risk and BD mood are limited, we adopt the following related baselines. Please refer to the Appendix C for a detailed explanation of the baselines.

BD Risk Detection: (i) BD detection (Sekulić et al., 2018): Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), (ii) MDD detection: DepRoBERTa (Poświata and Perełkiewicz, 2022), HAN-BERT (Zhang et al., 2022), SS3 (Burdisso et al., 2019), and AdaBoost (Paul et al., 2018), (iii) Large Language Models (LLMs) & Pre-trained Language Models (PLMs): ChatGPT[3] (Brown et al., 2020), MentaLLaMa (Yang et al., 2023b), BioClinicalBERT (Alsentzer et al., 2019), and PHS-BERT (Naseem et al., 2022b) and (iv) Mental Status Detection: UNSL (Loyola et al., 2021), PHASE (Sawhney et al., 2021), STATENet (Sawhney et al., 2020), PsyEx (Chen et al., 2023),

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

Table 4: Performance comparisons of the proposed model and baselines for BD risk detection, with the average results over 5-fold cross-validation.

Type	Model	Prec. ↑	Rec. ↑	F1 ↑
BD Detection	**SVM**	0.439	0.430	0.434
	LR	0.443	0.422	0.432
	RF	0.526	0.351	0.421
MDD Detection	**DepRoBERTa**	0.516	0.337	0.408
	HAN-BERT	0.531	0.490	0.510
	SS3	0.405	0.570	0.474
	AdaBoost	0.527	0.498	0.512
LLMs & PLMs	**ChatGPT**	*0.680*	0.070	0.130
	MentaLLaMa	**0.790**	0.110	0.190
	BioClinicalBERT	0.479	0.396	0.434
	PHS-BERT	0.536	0.264	0.353
Mental Status Detection	**UNSL**	0.507	0.466	0.486
	PHASE	0.398	0.532	0.457
	STATENet	0.443	0.572	0.499
	PsyEx	0.563	0.575	*0.569*
	UoS	0.479	0.495	0.487
	BD2SU	0.522	0.544	0.533
	Ours (STL)	0.482	*0.593*	0.532
	Ours (MTL)	0.540	**0.621**	**0.578**

Table 5: Comparison of performance between baselines and the proposed model for BD mood level and variation prediction.

Task		Mood Level		Mood Variation	
Type	**Model**	**MAE ↓**	**MDAE ↓**	**MAE ↓**	**MDAE ↓**
Sentiment Analysis	**BERT**	0.907	0.615	**0.895**	0.795
	EmoNet	0.857	0.709	0.934	0.630
	GoEmotions	0.804	0.698	*0.918*	**0.625**
	XLNet	0.838	0.631	0.980	0.713
	SBERT	0.757	0.604	0.933	*0.649*
Time-Series Detection	**UoS**	0.840	0.573	1.018	0.723
	BD2SU	0.733	*0.557*	0.944	0.651
	Ours (STL)	*0.705*	*0.525*	0.954	0.702
	Ours (MTL)	**0.701**	**0.501**	0.952	0.696

UoS (Azim et al., 2022), and BD2SU (Lee et al., 2023).

BD Mood Level & Variation Prediction: (i) Sentiment analysis : BERT (Devlin et al., 2019), EmoNet (Abdul-Mageed and Ungar, 2017), GoEmotions (Demszky et al., 2020), XLNet (Yang et al., 2019), and SBERT (Reimers and Gurevych, 2019) and (ii) Time-Series detection: UoS (Azim et al., 2022) and BD2SU (Lee et al., 2023).

5.3 Experiment Results

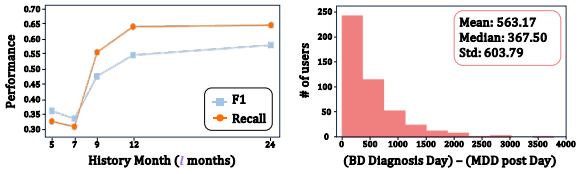
BD Risk Detection: Table 4 shows the experiment results for identifying the BD risk. Overall, the proposed model and most other time-series models outperform non-contextual models. Our proposed model also surpasses other sequential models, achieving an F1 score of 57.8%, mostly attributed to the dynamic mood-aware attention layer, which can effectively generate temporal contextual features by gaining deeper insights into the author's historical mood. Specifically, LLMs (Brown et al., 2020; Yang et al., 2023b) exhibited inferior performance compared to the proposed model. Previous studies on mental health analysis using LLMs typically involved simple binary classification to identify the current status of posts (Yang et al., 2023a). In contrast, our study focuses on predicting future BD risk based on historical posts. This suggests that LLMs still exhibit a notable gap due to a lack of mental health-specific knowledge and inconsistencies depending on prompting strategies (Yang et al., 2023a). Consequently, this poses a significant challenge in using LLMs for mental health analysis.

Table 6: Results of the ablation study on the proposed model components.

Model Components	Loss	Prec. ↑	Rec. ↑	F1 ↑
Temp. Conv.	L_r	**0.551**	0.409	0.469
+ Mood Att.	L_e	0.482	0.593	0.532
+ Mood Level	$L_r + L_m$	0.452	0.489	0.470
+ Mood Variation	$L_r + L_v$	0.500	0.669	0.572
+ Mood & Variation	$L_r + L_m + L_v$	0.524	**0.720**	**0.607**
+ Uncertainty (Ours)	L_T	0.540	*0.621*	*0.578*

BD Mood Level & Variation Prediction: Table 5 summarizes the Mean Absolute Error (MAE) and Median Absolute Error (MDAE) for the BD mood level and variation prediction. Our model exhibits more precise predictions than other baselines for the BD mood level prediction task, achieving an MAE of 0.701. The performance of the pretrained SBERT model surpasses other state-of-the-art semantic analysis models, which implies that SBERT can generate sentence embeddings with semantically significant information (Reimers and Gurevych, 2019). However, despite the high performance in predicting BD mood level, the results for predicting mood variation are relatively moderate and comparable. Our analysis reveals that there is a trade-off between these two tasks, and enhancing the performance of both tasks is future work.

Multi-task Learning: Notably, the performance of the proposed model using multi-task learning (MTL) is superior to employing single-task learning (STL) in both tasks. We believe that capturing BD mood is useful in identifying BD risk, and hence, jointly learning both tasks can improve performance by facilitating the transfer of informative representations and parameters among the tasks.



(a) (b)

Figure 3: (a) Performance of the proposed model by varying observational period l (from 5 to 24 months). (b) Distributions of diagnosis shift duration between the MDD diagnosis and BD diagnosis

5.3.1 Ablation Study

We perform an ablation study to assess the effectiveness of each component and multi-task learning of the proposed model, as shown in Table 6. We begin with the base model, where only the Temporal Convolution layer is applied, which does not involve multi-task learning. By applying the Dynamic Mood-Aware Attention layer, we find drastic improvements in F1 scores (46.9% $\rightarrow$ 53.2%), which emphasizes the importance of retaining temporal information in predicting a user's risk of BD. Our findings also indicate that jointly learning the mood variation task outperforms learning the mood level task alone, but considering both information together is the most effective. This highlights that acquiring both the absolute mood level and the relative mood variation is helpful in monitoring BD risk, aligning with previous clinical research (Ortiz et al., 2018). Notably, applying uncertainty weight parameters θ decreases performance significantly in recall but increases precision. This suggests that θ effectively manages the task weights during training, particularly considering their distinct levels of granularity, i.e., post-level for BD mood level prediction vs. user-level for BD risk detection. This careful balance prevents any individual task from dominating the overall objective function, ultimately enhancing the overall performance.

5.3.2 Observational and Predictable Periods

We examine how the observation period (l months) influences the prediction of future BD risk in the next period (m months). Since we found that the average diagnosis shift duration between the MDD diagnosis and BD diagnosis is 563.17 days (Std = 603.79) in Figure 3b, we set m at 24 months, which allows fair comparisons that are not significantly affected by fluctuations in m. Figure 3a presents the F1 and recall for the proposed model over 24 months, with l ranging from 5 to 24 months. Notably, the model performance improves with a more extended training history, which is associated with the recurrent mood swings exhibited by BD patients, necessitating a substantial duration to capture their patterns (Egeland et al., 2012). Interestingly, these findings do not align with the earlier study that proposed a 6-month observation period for suicide risk prediction of BD patients due to their rapid mood fluctuations (Lee et al., 2023). This differentiation can be interpreted as different observation periods that are required depending on the objectives of tasks.

5.3.3 Interpretability of the Model

To demonstrate the interpretability of the proposed model, we show an example sequence u_5 where our model performs better than other models. In particular, we analyze how the model assigns the attention weights a_t^i from the Dynamic Mood-Aware Attention layer to each post over time in predicting BD risk. Note that we compare the proposed model adopting multi-task learning and single-task learning (i.e., 'ours/MTL' vs. 'ours/STL' in Table 6) to demonstrate the benefits of joint learning. As shown in Figure 4, both models tend to have higher attention scores when the absolute mood level y_m is elevated, such as p_1^5 and p_{16}^5. This aligns with the fact that the presence of manic symptoms is a key factor for BD diagnosis (American Psychiatric Association et al., 2013). Interestingly, in the case of 'ours/STL', lower attention scores are assigned to $p_2^5 - p_4^5$ due to their negative mood level y_m, whereas 'ours/MTL' allocates more attention to these instances due to their relatively higher mood variation y_v. This implies that the 'ours/MTL' model is effective at identifying BD risk by not only assessing mood levels but also by recognizing mood swings, which are recognized as crucial indicators for diagnosis (Ortiz et al., 2018). We believe the proposed model with interpretability (such as Figure 4) has the potential to screen and detect individuals with BD on social media, allowing for early clinical intervention.

5.3.4 Error Analysis

We further analyze some typical errors during our experiments that can be used for future performance improvement.

Mood Swing Detection in *MDD* $\rightarrow$ *BD*: For the *MDD* $\rightarrow$ *BD* group, we find that users who failed to predict exhibited a more monotonous mood swing (average = 0.728) compared to those who

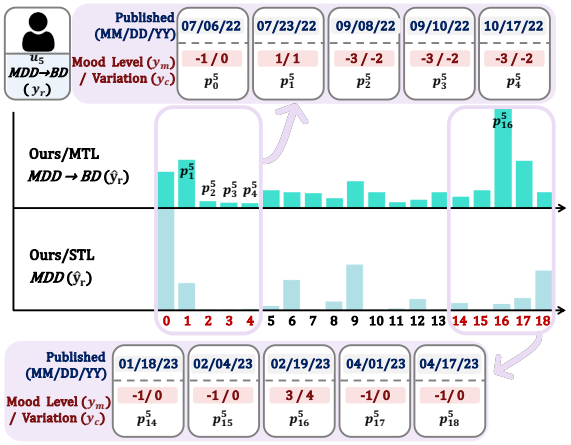


Figure 4: A case study illustrating how the model assigns attention weights a_t^i. Such an analysis can offer interpretability of the proposed model.

succeeded (average = 0.788). In other words, the standard deviation of the mood swing was confirmed to be smaller in error cases. This indicates that our model emphasizes capturing mood swing transitions in future BD risk detection. In future work, we will consider incorporating additional features related to mood swing patterns, such as intensity variations, or sentiment analysis, to enrich the model's understanding.

Post Volume Impact in *MDD*: In the case of *MDD*, the number of posts emerged as a significant variable. Through statistical analysis, we find that most users who successfully predicted had an average of 7.95 posts, while those who failed had an average of 5.15 posts. This is likely due to the higher volume of posts, suggesting that more information proves beneficial for accurate predictions. For future work, we will improve our sequence model to better capture the temporal dynamics of posts and their impact on predictions.

6 Concluding Remarks

Our research presented a multi-task learning model for BD risk detection among users misdiagnosed as MDD, leveraging a clinically validated novel dataset, *BD-Risk*. We demonstrated that jointly learning BD mood information can enhance the performance of BD risk detection. The dynamic mood-aware attention scores offer insights into BD mood's impact on future risk, potentially aiding clinicians in interventions for those who are underrepresented in a clinical setting, such as minorities or patients with a lack of insight.

Limitations

Evaluating moods on social media is subjective, and researchers may interpret the analysis of this paper differently (Keilp et al., 2012). Despite our careful annotation, there is a chance of including unreliable data if users misunderstand their diagnoses. Detailed explanations are described in the § 3.2 *Data Filtering*. Additionally, the effectiveness of using social media data for predicting mental health can be limited in certain clinical settings (Ernala et al., 2019). However, we believe our proposed model can help psychiatrists understand how their patients are doing in their daily lives outside of the hospital.

Ethics Statement

Our approach to addressing ethical concerns in this study includes two key aspects: (i) safeguarding the privacy of Reddit users and (ii) preventing any potentially harmful uses of the dataset we propose. We adhere to the guidelines in Reddit's privacy policy[4] and widely accepted social media research ethics policies, which permit researchers to utilize user data without explicit consent as long as anonymity is preserved (Benton et al., 2017; Williams et al., 2017).

It is important to note that we have not collected any metadata that could be used to identify the authors of the content. Moreover, all content undergoes a thorough manual review to remove personally identifiable information and mask any named entities. Most significantly, we ensure the responsible use of data while maintaining anonymity. The dataset will only be shared with researchers who commit to ethical principles. Our study obtained approval from the Institutional Review Board (IRB) (SKKU2022-11-038)).

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C2007625), and by the MSIT (Ministry of Science and ICT), Korea, under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-RS-2023-00259497) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

[4] https://www.reddit.com/policies/privacy-policy

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

DSMTF American Psychiatric Association, American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *IEEE Intelligent Systems*, 38:2.

Jules Angst, Jean-Michel Azorin, Charles L Bowden, Giulio Perugi, Eduard Vieta, Alex Gamma, Allan H Young, BRIDGE Study Group, et al. 2011. Prevalence and characteristics of undiagnosed bipolar disorders in patients with a major depressive episode: the bridge study. *Archives of general psychiatry*, 68(8):791–799.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 213–218.

Ross J Baldessarini, Eduard Vieta, Joseph R Calabrese, Mauricio Tohen, and Charles L Bowden. 2010. Bipolar depression: overview and commentary. *Harvard review of psychiatry*, 18(3):143–157.

Elizabeth D Ballard, Cristan A Farmer, Bridget Shovestul, Jennifer Vande Voort, Rodrigo Machado-Vieira, Lawrence Park, Kathleen R Merikangas, and Carlos A Zarate Jr. 2020. Symptom trajectories in the months before and after a suicide attempt in individuals with bipolar disorder: A step-bd study. *Bipolar disorders*, 22(3):245–254.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Charles L Bowden, Vivek Singh, P Thompson, JM Gonzalez, MM Katz, M Dahl, Thomas J Prihoda, and X Chang. 2007. Development of the bipolar inventory of symptoms scale. *Acta Psychiatrica Scandinavica*, 116(3):189–194.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sergio G Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. 2019. Unsl at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media. In *CLEF (Working Notes)*.

Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084, Singapore. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International conference on machine learning*, pages 2067–2075. PMLR.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. 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.

John Daveney, Maria Panagioti, Waquas Waheed, and Aneez Esmail. 2019. Unrecognized bipolar disorder in patients with depression managed in primary care: a systematic review and meta-analysis. *General hospital psychiatry*, 58:71–76.

Richard J Davidson, Klaus R Sherer, and H Hill Goldsmith. 2009. *Handbook of affective sciences*. Oxford University Press.

Jorge Renner Cardoso de Almeida and Mary Louise Phillips. 2013. Distinguishing between unipolar depression and bipolar depression: current and future clinical and neuroimaging perspectives. *Biological psychiatry*, 73(2):111–118.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Janice A Egeland, Jean Endicott, Abram M Hostetter, Cleona R Allen, David L Pauls, and Jon A Shaw. 2012. A 16-year prospective study of prodromal features prior to bpi onset in well amish children. *Journal of affective disorders*, 142(1-3):186–192.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Bruno Etain, Mohamed Lajnef, Frank Bellivier, Flavie Mathieu, Aurélie Raust, Barbara Cochet, Sébastien Gard, M Katia, Jean-Pierre Kahn, Orly Elgrabli, et al. 2012. Clinical expression of bipolar disorder type i as a function of age and polarity at onset: convergent findings in samples from france and the united states. *The Journal of clinical psychiatry*, 73(4):2757.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73.

Xiaobo Guo, Yaojia Sun, and Soroush Vosoughi. 2021. Emotion-based modeling of mental disorders on social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 8–16.

Daisy Harvey, Fiona Lobban, Paul Rayson, Aaron Warner, Steven Jones, et al. 2022. Natural language processing methods and bipolar disorder: Scoping review. *JMIR mental health*, 9(4):e35928.

Ya-Han Hu, Kuanchin Chen, I-Chiu Chang, and Cheng-Che Shen. 2020. Critical predictors for the early detection of conversion from unipolar major depressive disorder to bipolar disorder: nationwide population-based retrospective cohort study. *JMIR medical informatics*, 8(4):e14278.

Jena D. Hwang and Kristy Hollingshead. 2016. Crazy mad nutters: The language of mental health. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 52–62, San Diego, CA, USA. Association for Computational Linguistics.

Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven JM Jones. 2021. Understanding who uses reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 1–14.

S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. 2022. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 331–338.

John G Keilp, Michael F Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K Burke, Hanga Galfalvy, Maria A Oquendo, and J John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of affective disorders*, 140(1):75–81.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Kamyar Keramatian, Jairo V Pinto, Ayal Schaffer, Verinder Sharma, Serge Beaulieu, Sagar V Parikh, and Lakshmi N Yatham. 2022. Clinical and demographic factors associated with delayed diagnosis of bipolar disorder: data from health outcomes and patient evaluations in bipolar disorder (hope-bd) study. *Journal of Affective Disorders*, 296:506–513.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, 10(1):1–6.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Towards suicide prevention from bipolar disorder with temporal symptom-aware multitask learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4357–4369, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Usha Lokala, Aseem Srivastava, Triyasha Ghosh Dastidar, Tanmoy Chakraborty, Md Shad Akhtar, Maryam Panahiazar, and Amit Sheth. 2022. A computational approach to understand mental health from reddit: knowledge-aware multitask learning framework. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 640–650.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th International Conference of the CLEF Association, CLEF*, pages 1–20.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 340–357. Springer.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Juan Martín Loyola, Sergio Burdisso, Horacio Thompson, Leticia C Cagnina, and Marcelo Errecalde. 2021. Unsl at erisk 2021: A comparison of three early alert policies for early risk detection. In *CLEF (Working Notes)*, pages 992–1021.

John Lu, Sumati Sridhar, Ritika Pandey, Mohammad Al Hasan, and Georege Mohler. 2019. Investigate transitions into drug addiction through text mining of reddit data. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2367–2375.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 825–832.

Giovanna Menardi and Nicola Torelli. 2014. Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28(1):92–122.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human.

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

Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022b. Benchmarking for public health surveillance tasks on social media with a domain-specific pretrained language model. *NLP-Power 2022*, page 22.

Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on computational research in phonetics, phonology, and morphology*, pages 82–86.

Abigail Ortiz, Kamil Bradler, Julie Garnham, Claire Slaney, and Martin Alda. 2015. Nonlinear dynamics of mood regulation in bipolar disorder. *Bipolar disorders*, 17(2):139–149.

Abigail Ortiz, Kamil Bradler, and Arend Hintze. 2018. Episode forecasting in bipolar disorder: Is energy better than mood? *Bipolar disorders*, 20(5):470–476.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, pages 864–887.

Ives C Passos, Pedro L Ballester, Rodrigo C Barros, Diego Librenza-Garcia, Benson Mwangi, Boris Birmaher, Elisa Brietzke, Tomas Hajek, Carlos Lopez Jaramillo, Rodrigo B Mansur, et al. 2019. Machine learning and big data analytics in bipolar disorder: a position paper from the international society for bipolar disorders big data task force. *Bipolar Disorders*, 21(7):582–594.

Sayanta Paul, Sree Kalyani Jandhyala, and Tanmay Basu. 2018. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *CLEF (Working notes)*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Maurizio Pompili, Xenia Gonda, Gianluca Serafini, Marco Innamorati, Leo Sher, Mario Amore, Zoltan Rihmer, and Paolo Girardi. 2013. Epidemiology of suicide in bipolar disorders: a systematic review of the literature. *Bipolar disorders*, 15(5):457–490.

Rafał Poświata and Michał Perełkiewicz. 2022. Opi@ lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.

Aswin Ratheesh, Christopher Davey, Sarah Hetrick, Mario Alvarez-Jimenez, Catherine Voutier, Andreas Bechdolf, Patrick D McGorry, Jan Scott, Michael Berk, and Susan M Cotton. 2017. A systematic review and meta-analysis of prospective transition from major depression to bipolar disorder. *Acta Psychiatrica Scandinavica*, 135(4):273–284.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Zoltán Rihmer and Kitty Kiss. 2002. Bipolar disorders and suicidal behaviour. *Bipolar Disorders*, 4:21–25.

Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Shah. 2021. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 2415–2428.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7685–7697.

Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 72–78.

Junneng Shao, Zhongpeng Dai, Rongxin Zhu, Xinyi Wang, Shiwan Tao, Kun Bi, Shui Tian, Huan Wang, Yurong Sun, Zhijian Yao, et al. 2019. Early identification of bipolar from unipolar depression before manic episode: evidence from dynamic rfmri. *Bipolar disorders*, 21(8):774–784.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023a. 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.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Minjoo Yoo, Sangwon Lee, and Taehyun Ha. 2019. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Information processing & management*, 56(4):1565–1575.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022. Psychiatric scale guided risky post screening for early detection of depression. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5220–5226. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

A BD-Risk Dataset

For a comprehensive understanding of each category and example, please refer to Table 7 in the Appendix A, which we used for the annotator instruction.

A.1 Data Analysis

We analyze our dataset, *BD-Risk*, to understand the similarities and differences between *MDD* and *MDD* $\rightarrow$ *BD* groups.

A.1.1 User Behavior Differences

For a more precise data analysis, we compare the user behavior differences between *MDD* and *MDD* $\rightarrow$ *BD*, providing the word count and time interval

Table 7: The descriptions and examples of BD Mood Levels.

BD Mood	Description & Examples
3	Severe restlessness with "hallucinations" or "delusions". *"I hear thoughts that aren't mine, or voices."*
2	Excessive self-confidence and overwhelming plans. *"doing stupid stuff like opening car door while driving."*
1	High motivation, self-confidence, and a positive mood. *"I've gotten better before, I'll get better again."*
0	In a moderate and comfortable state of mind. *"I am improving, but I am unsure if it will be enough."*
-1	Feeling down, a decrease in self-confidence, and motivation. *"I feel stupid for turning it down."*
-2	Severe depression and inability to perform daily tasks. *"I don't have money to get back on my medications."*
-3	Experiencing extreme anxiety and having suicidal thoughts. *"My depression has made me suicidal."*

Table 8: Results of the differences between *MDD* and *MDD* $\rightarrow$ *BD* groups based on the word count and the time interval between posts. We report the average of results (*avg.*). * indicates that the result is statistically significant ($p < 0.05$) under the t-test.

	MDD	MDD $\rightarrow$ BD	t
avg. # Word Count	1163.02	1063.51	-2.90 *
avg. # Time Interval Between Posts	73.33	66.04	1.64

results in Table 8 using the t-test. For the average word count, we find that *MDD* had an average of 1163.02 words, while *MDD* $\rightarrow$ *BD* had an average of 1063.51 words. Furthermore, we calculate the average time interval between sequential posts of each user, as we annotated users' posts based on mood levels (§ 3.2 *Annotation Process*). Please refer to the distribution of the duration of diagnosis shift from MDD to BD diagnoses, illustrated in Figure 3b.

A.1.2 Keywords Usage Differences

To identify the keywords usage differences between *MDD* and *MDD* $\rightarrow$ *BD*, we apply the odds ratio (Lu et al., 2019), which indicates the likelihood of an event occurring in a group compared to another. The odds ratio $OR(g, w)$ is calculated for each 1,146 words w extracted from TF-IDF between the groups $g \in \{MDD, MDD{\rightarrow}BD\}$ as follows:

$$OR(g, w) = \frac{Freq(g, w) * \neg Freq(g, w)}{Freq(\neg g, w) * \neg Freq(g, w)} \quad (15)$$

where $Freq(g, w)$ is the number of posts that include w in g, and $\neg Freq(g, w)$ is the number of posts that do not include w. Table 9 shows distinct keywords extracted for each group based on odds ratio values exceeding 1.5 (Nguyen and Cornips, 2016), e.g., 'appetite' is approximately 1.5 times more likely shown in *MDD* $\rightarrow$ *BD* than in *MDD*. We find that *MDD* contains a higher prevalence of words related to everyday life and risk factors (e.g., crisis and ignored), suggesting consistent low mood variability aligning with the clinical studies (Ortiz et al., 2015). On the other hand, *MDD* $\rightarrow$ *BD* includes words commonly associated with medication (e.g., abilify and ssri) (Yoo et al., 2019) or symptoms widely reported by BD patients in clinical settings (e.g., appetite and abused) (Bowden et al., 2007). These linguistic differences highlight the utility of social media data that can be used to analyze both conditions.

Table 9: Keywords usage differences between *MDD* and *MDD* $\rightarrow$ *BD* groups based on the odds ratio results.

Group	Words (Odds Ratio >1.5)
MDD	depression, crisis, covid, friends, family, pushing, ptsd, cared, anymore just, ignored, joke, anxiety, hates, exams
MDD $\rightarrow$ BD	abilify, appetite, surgery, effexor, unemployed, dose, ssri, psych, zoloft, experienced, abused, anybody

A.1.3 Mood Level Differences

We compare the two groups with their annotation data and LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001) results using the t-test. Table 10 indicates that *MDD* $\rightarrow$ *BD* exhibits notably higher levels of manic moods (0 to 3), while *MDD* shows significantly higher levels of depressive moods (-3 to -1). Additionally, similar trends are observed in LIWC, with *negemo* and *sad* scores being greater in *MDD* than *MDD* $\rightarrow$ *BD*. Furthermore, *MDD* $\rightarrow$ *BD* displays elevated scores in the *sexual*, *ingest*, and *death*, which are linked to prominent symptoms of BD, such as increased sexuality, eating disorders, and suicide ideation (Ballard et al., 2020).

B Experimental Settings

We adopted a stratified 5-fold cross-validation. We split the data set into train and test sets at an 8:2 ratio, ensuring that users in the test set were entirely separate and did not overlap with those in the training set. We fine-tuned the hyperparameters based on each model's highest F1 score from the cross-validation set. We implemented all methods using PyTorch 1.12 and optimized with

Table 10: Differences between the *MDD* → *BD* and *MDD* (* : p-value < 0.05, ** : p-value < 0.005).

BD Mood	t	LIWC	t
3	4.62 **	**posemo**	-1.10
2	7.59 **	**negemo**	-2.15 *
1	1.85	**anger**	1.44
0	0.23	**sad**	-3.96 **
-1	-1.43	**affect**	-2.18 *
-2	-1.69	**sexual**	2.20 *
-3	-3.68 **	**ingest**	3.50 **
		death	0.83

the AdamW (Loshchilov and Hutter, 2018), Exponential Learning rate Scheduler with gamma 0.1, $lr = 1e-4$, and with a batch size of 32. We trained the model on a GeForce RTX 3090 Ti GPU for 50 epochs and applied early stopping with the patience of 7 epochs. To solve the imbalanced data issue, random oversampling (Menardi and Torelli, 2014) is applied.

C Baseline Models

For comprehensive performance comparisons, we evaluate various baselines. As studies on identifying (i) BD risk and (ii) BD mood level & variation are limited, we adopt the following related baselines.

C.1 BD Risk Detection

C.1.1 BD Detection

Due to limited studies for detecting BD diagnosis, three different machine learning methods are used as our baseline models from; (i) *Support Vector Machine (SVM)*, (ii) *Logistic Regression (LR)*, and (iii) *Random Forest (RF)* with different features (i.e., LIWC, Empath, and TF-IDF) (Sekulić et al., 2018).

C.1.2 MDD Detection

As our goal is identifying BD risk in individuals initially misdiagnosed with MDD, we employ baseline models aiming to investigate early warning signals of depression issues on social media data.

- **DepRoBERTa** (Poświata and Perełkiewicz, 2022)[5]: *DepRoBERTa* is the winning model at the DepSign-LT-EDI@ACL_2022 challenge (Kayalvizhi et al., 2022), focusing on classifying depression signs through social media data. This model was pre-trained on subreddits, such as *r/depression* and *r/SuicideWatch*, and was further pre-trained on RoBERTa (Liu et al., 2019).

- **HAN-BERT** (Zhang et al., 2022): *HAN-BERT* is a BERT (Devlin et al., 2019) based model emphasizing the significance of early identification of risky posts in the context of clinical depression detection. The model employs a hierarchical attentional network to select critical content and make predictions.

- **SS3** (Burdisso et al., 2019)[6]: *SS3* is a supervised machine learning model for early risk detection by learning specific parameters related to three critical aspects: *Smoothness*, *Significance*, and *Sanction*. This is the best-performing model at CLEF eRisk 2019 challenge (Losada et al., 2019).

- **AdaBoost** (Paul et al., 2018): Paul et al. (2018) utilized various machine learning methods employing a simple bag of words model to detect early signs of depression at CLEF eRisk 2018 (Losada et al., 2018); (i) *AdaBoost*, (ii) *LR*, (iii) *SVM*, and (iv) *RF*. Please refer to the Appendix C.1.1 *BD Detection* for more details.

C.1.3 LLMs & PLMs

While recent advancements in utilizing LLMs have showcased robust capabilities in general language processing, numerous studies have highlighted the unreliability and inconsistency of LLMs when applied to mental health analysis (Amin et al., 2023; Yang et al., 2023a; Lamichhane, 2023). To examine the performance of LLMs as baselines, we also incorporate experiments involving both well-known LLMs and PLMs to ensure a more precise evaluation.

- **ChatGPT** (gpt-3.5-turbo) (Brown et al., 2020)[7]: Developed by OpenAI, *ChatGPT* is a text generation model specialized in generating conversational text through reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020). *ChatGPT* is based on the 175 billion parameters version of InstructGPT (Ouyang et al., 2022).

[5] https://huggingface.co/rafalposwiata/deproberta-large-v1

[6] https://pyss3.readthedocs.io/en/latest/

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

- **MentaLLaMa** (Yang et al., 2023b)[8]: *MentaLLaMA* is the open-source LLM fine-tuned with 105K mental health-related social media data for interpretable mental health analysis. We test on MentaLLaMA with 7 billion (MentaLLaMA-chat-7B) parameters.

- **BioClinicalBERT** (Alsentzer et al., 2019)[9]: *BioClinicalBERT* is a pre-trained BERT (Devlin et al., 2019) model specifically for the clinical domain, which was trained on MIMIC III, an electronic health records database.

- **PHS-BERT** (Naseem et al., 2022b)[10]: *PHS-BERT* is a transformer-based BERT (Devlin et al., 2019) model pre-trained to detect tasks associated with public health surveillance (PHS) on social media, such as health-related tweets.

C.1.4 Time Series Mental Status Detection

We employ baseline models from CLEF's eRisk Challenges 2021 (Parapar et al., 2021), aiming to investigate early warning signals of mental status-related issues (e.g., pathological gambling and self-harm) on social media data. Furthermore, as using longitudinal social media data is vital in finding an individual who may develop BD after an initial diagnosis of MDD; we utilize time series detection models used in mental disorder detection as our baselines.

- **UNSL** (Loyola et al., 2021): To find the signal of pathological gambling, *UNSL* applies SVM-RBF using a Bag-of-Words representation by employing 4 grams at the character level and TF-IDF.

- **PHASE** (Sawhney et al., 2021): *PHASE* is a time-sensitive transformer-based model to generate a temporal context for predicting a user's suicidality. EmoNet (Abdul-Mageed and Ungar, 2017) is used for encoding tweets.

- **STATENet** (Sawhney et al., 2020): *STATENet* is a transformer-based model highlighting the importance of emotional and temporal context for assessing suicide risk. This model employs two distinct embeddings to capture information from different time periods; we leverage Sentence-BERT (Reimers and Gurevych, 2019) to create embeddings for the most recent posts, while historical period embeddings are generated using EmoNet (Abdul-Mageed and Ungar, 2017).

- **PsyEx** (Chen et al., 2023): *PsyEx* simultaneously identifies multiple mental diseases by employing a two-stream architecture that processes both text and symptom features from social media data. This approach combines the strengths of both modalities, resulting in enhanced detection performance.

- **UoS** (Azim et al., 2022): *UoS* is the best-performing model at CLPsych 2022 shared task (Zirikly et al., 2019), employing a multi-task learning approach using a transformer encoder and Bi-LSTM to predict (i) changes in a user's mood and (ii) their level of suicidal risk.

- **Bipolar To Suicide (BD2SU)** (Lee et al., 2023): The *BD2SU* is a multi-task model for (i) BD symptom identification and (ii) future suicidality prediction. This model uses the attention layer with learnable parameters for considering temporal information.

C.2 BD Mood Level & Variation Prediction

C.2.1 Sentiment Analysis

To predict BD mood level and variation, state-of-the-art models of sentiment analysis are used, which is a method for identifying emotions expressed in text.

- **BERT** (Devlin et al., 2019)[11]: Fintuned *BERT* has shown great performance across various text benchmarks, including sentiment analysis, by adopting masked language modeling and next-sentence prediction for training.

- **EmoNet** (Abdul-Mageed and Ungar, 2017): *EmoNet* is applied a Gated Recurrent Neural Networks (Chung et al., 2015) that modifies LSTM with a reset gate, an updated state, and a hidden unit using 250m tweets with 24 emotions consisting of 665 emotion hashtags annotated labels.

[8] https://huggingface.co/klyang/MentaLLaMA-chat-7B

[9] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

[10] https://huggingface.co/publichealthsurveillance/PHS-BERT

[11] https://huggingface.co/bert-base-uncased

- **GoEmotions** (Demszky et al., 2020)[12]: A pretrained RoBERTa *GoEmotions* is finetuned based on the GoEmotions dataset, including 58k English Reddit comments, labeled by 80 human annotators across 28 emotion categories.
- **XLNet** (Yang et al., 2019)[13]: *XLNet* is a generalized autoregressive pretraining method that outperforms BERT on sentiment analysis tasks.
- **Sentence-BERT (SBERT)** (Reimers and Gurevych, 2019): *SBERT* is a modified version of the pre-trained *BERT*, which employs siamese and triplet network structures to generate sentence embeddings with semantic meaning.

C.2.2 Time Series Mental Status Detection

UoS (Azim et al., 2022) and ***BD2SU*** (Lee et al., 2023) both adopt multi-task learning to predict sequence posts. Please refer to the Appendix C.1.4 *Time Series Mental Status Detection* for more details.

[12] https://huggingface.co/SamLowe/roberta-base-go_emotions

[13] https://huggingface.co/xlnet-base-cased