PAT: A Personality-Driven Augmentation and Transfer Learning Framework for Depression Detection on Social Media

Anonymous ACL submission

Abstract

Depression is a prevalent mental disorder affecting millions worldwide, with early detection crucial for effective intervention. While existing methods have achieved remarkable results in automated depression detection, they face two key limitations: (1) Weak optimization of post encoders: Relying solely on coarse user-level supervision signals prevents models from capturing depressive cues within individual posts; (2) Lack of interpretability: Current frameworks cannot substantiate their outputs with granular evidence, which undermines their trustworthiness. To address these challenges, we propose PAT, a **P**ersonality-driven **A**ugmentation and **T**ransfer learning framework. PAT first optimizes the post encoder on the post-level depression detection and then transfers it to the user-level task. To fully utilize the scarce post-level data and enhance encoding performance, PAT also introduces personality-driven augmentation and fine-grained contrastive learning. Extensive experiments demonstrate that PAT significantly outperforms existing baselines. Moreover, PAT provides comprehensive interpretability by delivering user-level predictions, tracing post-level mental-state trajectories, and highlighting key symptoms, thereby offering valuable diagnostic evidence for clinical practice.

1 Introduction

Depression represents a prevalent and debilitating mental disorder in modern society. According to the World Health Organization (WHO)[1], depression affected approximately 332 million individuals globally in 2021, corresponding to 4.0% of the world's population. Its core symptoms include sadness, pessimism, self-dislike and other affective disturbances, which can lead to suicidal behavior in severe cases (Beck, 1996). Early detection and intervention are critical for mitigating disease progression and preventing these adverse outcomes.

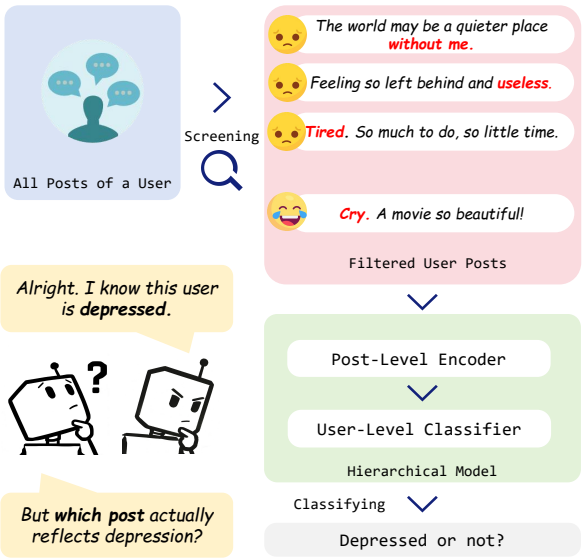


Figure 1: A typical Screen-and-Classify framework for social media user-level depression detection. The hierarchical model is trained solely on binary supervision (depressed or non-depressed), which prevents the post encoder from capturing fine-grained depressive cues in each post and thus limits its optimization. Moreover, the screening step may introduce false positives. For instance, the word "cry" in the fourth sentence of the figure expresses a positive emotion, misleading the model.

With the popularization of social networks, a growing number of users are willing to express their real emotions and personal feelings online, which provides extensive data for developing automated tools for depression detection (Tahir et al., 2025; Wang et al., 2024; Ding et al., 2020). Within this domain, the primary task is to predict a user's depressive status according to their posts, commonly termed user-level depression detection. A leading approach for this task is the Screen-and-Classify framework (Zhang et al., 2022; Liu et al., 2024; Wang et al., 2025), as illustrated in Figure 1. This framework comprises two steps: (1) Post

[1] https://www.who.int/news-room/fact-sheets/detail/depression

Screening: retrieving posts related to depression using clinical descriptions as queries; and (2) Hierarchical Classification: encoding the posts and aggregating embeddings for classification. However, this framework is optimized solely with simple user-level binary supervision, which fails to provide sufficient guidance for the post encoder, thus causing a performance bottleneck. Besides, these methods only output user-level predictions with limited explanations, lacking the evidence needed for clinical interpretation. Recent efforts have attempted to address this gap by leveraging Large Language Models (LLMs) for post-level information extraction and explanation generation (Lan et al., 2025; Zheng et al., 2024). While promising, LLM-based approaches face their own constraints: the risk of hallucination and high computational costs that hinder real-world deployment(Zhang et al., 2025; Ravenda et al., 2025).

To address these limitations, we propose the **P**ersonality-driven **A**ugmentation and **T**ransfer learning framework (PAT), whose overall structure is shown in Figure 2. PAT tackles the bottleneck in the post encoder through a two-phase approach: it first trains the encoder on a post-level detection task to learn precise depression features, and then leverages this optimized encoder in the user-level model for further training. However, the data scarcity problem poses a major challenge for post-level tasks. An important psychological study indicates that language serves as a latent behavioral manifestation of personality, and individuals with different personality traits exhibit significant linguistic differences in their expression (Koutsoumpis et al., 2022). Informed by this finding, we leverage a large language model for data augmentation driven by psychological personality theories including the Big Five (Costa and McCrae, 2008) and MBTI (Jung and Beebe, 2016), which is able to enrich data diversity and enable the model to adapt to diverse expressions. Besides, we design a corresponding fine-grained contrastive learning method to effectively utilize both the augmented and labeled data for representation learning. For the user-level model, we adopt an architecture similar to DeCapsNet (Liu et al., 2024), which incorporates a symptom attention module to inject medical knowledge and enable symptom-wise analysis. More importantly, PAT can provide multi-granular interpretability: it delivers user-level predictions, traces post-level mental-state trajectories, and highlights key symptoms ,thereby offering comprehensive diagnostic evidence. The contributions of this paper are threefold:

- We propose a novel transfer learning framework, PAT. Extensive experiments on multiple public benchmarks demonstrate that PAT significantly outperforms existing strong baselines and exhibits excellent generalization ability.
- We design a personality-driven data augmentation method with a corresponding fine-grained contrastive learning algorithm. Experiments demonstrate that these methods effectively optimize the post encoder's embedding space, significantly enhancing its ability to capture semantic patterns of depression.
- We build a comprehensive interpretable analysis system. Beyond user-level depression detection, PAT traces post-level mental-state trajectories and highlights key symptoms offering insights that enhance practical utility and trustworthiness.

2 Related Work

2.1 User-Level Depression Detection

User-Level Depression Prediction aims to assess depression risk using a user's social media posts. Early researchers explored diverse machine learning techniques(Ding et al., 2020; Vasha et al., 2023; Hossain et al., 2021). With the rapid advancement of deep learning, various neural networks have become the dominant approach for this task, such as Convolutional Neural Networks(Yates et al., 2017; Lin et al., 2020; Narayanan et al., 2022) and Recurrent Neural Networks(Kour and Gupta, 2022; Kour and Gupta, 2022; Gamaarachchige and Inkpen, 2019). Zhang et al. (2022) proposed a post screening method based on psychological scales and adopted a hierarchical model that sequentially conducts post encoding and user classifying, which has since become a mainstream approach in depression detection. Following this paradigm, Liu et al. (2024) introduced capsule networks to model features of different symptoms. Wang et al. (2025) further improved the screening method by making it trainable. However, these methods typically rely on user-level binary labels for the whole training. This makes it difficult for the models to obtain specific supervision signals at the post level.

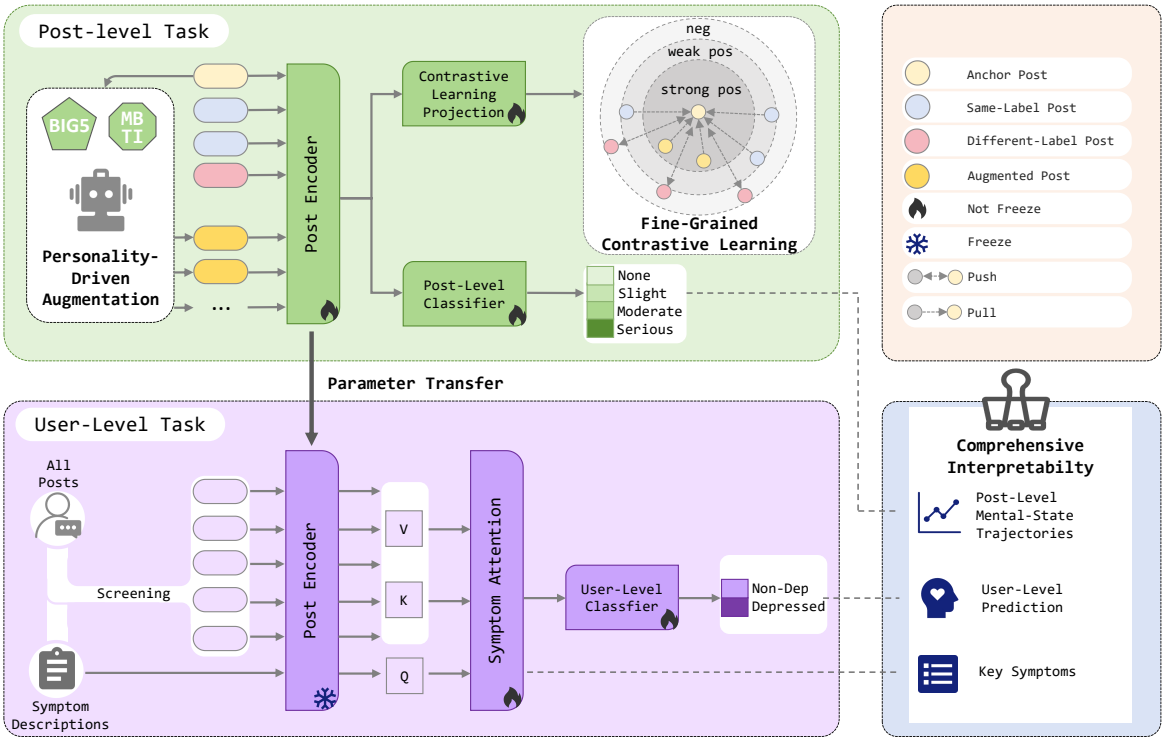


Figure 2: An overview of the PAT framework: it first performs post-level optimization and then transfers the post encoder to the user-level depression detection task. For post-level optimization, PAT introduces Personality-Driven Augmentation and novel Fine-grained Contrastive Learning. Unlike standard supervised contrastive learning, it distinguishes three data types derived from an anchor post: (1) **Augmented Posts**, generated under personality guidance, serving as strong positives; (2) **Same-Label Posts**, which share the ground-truth label, serving as weak positives; and (3) **Different-Label Posts**, which have different labels, serving as negatives.

Several research efforts have attempted to address this limitation. Pérez et al. (2023) designed a semantic retrieval framework that performs detection via K-nearest neighbor voting, which is not trainable. Nguyen et al. (2022) trained a separate questionnaire model with a semi-supervised dataset, which resulted in noisy labels. Ahmed et al. (2022) proposed a curriculum annotation method to enhance data quality. Agarwal et al. (2024) leveraged medical experts to annotate symptom-level scores, but this approach is costly and suffers from label sparsity because certain symptoms are rarely mentioned in personal posts due to privacy concerns.

2.2 Transfer Learning

The core of transfer learning is to leverage knowledge from a data-rich source domain to alleviate data scarcity in a target domain (Pan and Yang, 2009). In mental-health research, Ji et al. (2022) pre-trained a domain-specialized Mental-BERT on psychological texts. Wu et al. (2023) transferred knowledge from sentiment classification to depression detection via distillation, yielding promising results. These methods perform pre-training or transfer on domain-related corpora or auxiliary tasks, whose alignment with the target task remains suboptimal. In this work, we propose to use post-level depression detection as the source task and transfer the post encoder to the user-level depression detection model. This design is expected to achieve tighter knowledge transfer by directly bridging the gap between the two closely related levels of depression detection.

3 Methods

The architecture of the proposed PAT framework is shown in Figure 2. The framework follows a two-stage design, comprising a post-level task and a user-level task of depression detection. The post-level task acts as an auxiliary objective, designed to enhance the post encoder, whose parameters are subsequently transferred to the user-level task.

3.1 Post-Level Depression Detection

Problem Definition: The input is a single social media post denoted as p. The objective is

to predict a four-level depression-intensity label $l \in \{0, 1, 2, 3\}$, corresponding to the categories "None", "Slight", "Moderate" and "Serious" respectively.

3.1.1 Personality-Driven Augmentation

We employ psychological personality theories to guide a large language model (LLM) in generating augmented samples that reflect distinct personality tendencies, thereby mitigating the scarcity of post-level annotated data and thus enabling the encoder to adapt to diverse expressions. Specifically, we adopt two mainstream personality theories: (1) the Big Five: taking the high and low tendencies of each of the five traits (Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness), resulting in ten categories; (2) the MBTI (Jungian eight-function): directly using its eight cognitive-function dimensions including Extroverted, Introverted, Sensing, Intuitive, Thinking, Feeling, Judging, and Perceiving. We constrain the LLM to preserve the original depressive-symptom semantics of the input while accentuating the linguistic style associated with the specified personality trait. This approach achieves style-diverse and semantically controlled text augmentation. The detailed prompts are shown in Appendix A.1.

After data augmentation, we perform post-level model training. First, a pretrained encoder M_{post} encodes the post p into a text embedding $\boldsymbol{h}$. Then, we utilize a two-layer linear classifier to output the predicted label distribution $\hat{y}$. Let y denote the one-hot vector of the ground-truth label for p. The cross-entropy loss is computed as:

$$\mathcal{L}_{ce} = -y \cdot log(\hat{y}) \tag{1}$$

3.1.2 Fine-Grained Contrastive Learning

In addition to the classification objective, we introduce fine-grained contrastive learning to fully leverage augmented data for optimizing the embedding space. The core idea of contrastive learning is to pull positive (same-class) samples closer and push negative (different-class) samples apart. Our proposed fine-grained contrastive learning distinguishes two types of positive samples: strong positives (augmented posts) and weak positives (same-label posts), as illustrated in Figure 2.

Specifically, for a post embedding $\boldsymbol{h}_i$ in a data batch, we transform it via a projection head Φ into a contrastive embedding $\boldsymbol{z}_i$. Correspondingly, we obtain the set of strong-positive contrastive embeddings $\tilde{Z} = \{\tilde{\boldsymbol{z}}_1, ..., \tilde{\boldsymbol{z}}_m\}$ and the set of weak-positive embeddings $\dot{Z} = \{\dot{\boldsymbol{z}}_1, ..., \dot{\boldsymbol{z}}_t\}$. The fine-grained contrastive loss is formulated as:

$$\mathcal{L}_{cl} = \frac{1}{B} \sum_{i \in I} \frac{-1}{|\Gamma(i)|} \sum_{j \in \Gamma(i)} \log \frac{w_{ij} \cdot e^{sim(\boldsymbol{z}_i, \boldsymbol{z}_j) \setminus \tau}}{\sum_{k \in I \setminus i} e^{sim(\boldsymbol{z}_i, \boldsymbol{z}_k) \setminus \tau}} \tag{2}$$

where B denotes the batch size, $I = \{1, 2, ...B\}$ indexes all samples in a batch, $I \setminus i$ denotes the index set of the batch excluding the i-th sample, $\Gamma(i)$ represents the set of indices of all positive samples (strong or weak) for the i-th sample, $sim(\boldsymbol{z}_i, \boldsymbol{z}_j)$ is the cosine similarity between the anchor sample $\boldsymbol{z}_i$ and one of its positive samples $\boldsymbol{z}_j$. τ is the temperature hyper-parameter. w_{ij} is the weight that is set to 1 for strong positive samples, and dynamically varies between w_{low} and 1 for weak positive samples, depending on its distance to the anchor sample. We define the distance between samples as follows:

$$d_{ij} = 1 - sim(\boldsymbol{z}_i, \boldsymbol{z}_j) \tag{3}$$

The average distance from strong positive samples $\tilde{Z}$ to the anchor sample $\boldsymbol{z}_i$ is calculated as follows:

$$\overline{d}_i = \frac{1}{|\tilde{Z}|} \sum_{\tilde{\boldsymbol{z}} \in \tilde{Z}} (1 - sim(\boldsymbol{z}_i, \tilde{\boldsymbol{z}})) \tag{4}$$

For a weak positive sample $\boldsymbol{z}_j$, its weight decays exponentially with the distance:

$$w_{ij} = \begin{cases} w_{low} + (1 - w_{low}) e^{-\gamma \frac{d_{ij} - \overline{d}_i}{d_{ij}}} & \text{if } d_{ij} > \overline{d}_i \\ 1 & \text{if } d_{ij} \leq \overline{d}_i \end{cases} \tag{5}$$

where γ is a hyper-parameter and w_{low} is the decay lower bound.

Finally, let the weighted coefficient for $\mathcal{L}_{cl}$ be α, and the overall loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \alpha \mathcal{L}_{cl} \tag{6}$$

3.2 User-Level Depression Detection

Problem Definition: The input is a set of social media posts from a user, denoted as $P = \{p_1, ..., p_n\}$ where p_i is the i-th post. The objective is to predict a binary depression label $l \in \{0, 1\}$, with 0 indicating non-depressed and 1 indicating depressed.

3.2.1 Post Screening

Given a set of user posts $P = \{p_1, \ldots, p_n\}$, we adopt the screening approach of Zhang et al. (2022), which leverages symptom descriptions from clinical scales to identify depressive content. Specifically, we derive 21 symptom descriptions from the Beck Depression Inventory-II (BDI-II) (Beck, 1996), as provided in Appendix A.4. We employ the Sentence-BERT (Reimers and Gurevych, 2019) to obtain sentence embeddings for all posts and symptom descriptions, denoted as $\boldsymbol{P} = \{\boldsymbol{p_1}, ..., \boldsymbol{p_n}\}$ and $\boldsymbol{S} = \{\boldsymbol{s_1}, ..., \boldsymbol{s_{21}}\}$ respectively. The depression risk score for each post p_i is then calculated as follows:

$$r_i = \sum_{\boldsymbol{s} \in \boldsymbol{S}} \frac{\boldsymbol{p}_i^T \boldsymbol{s}}{||\boldsymbol{p}_i|| \cdot ||\boldsymbol{s}||} \quad i = 1, 2, ..., n \tag{7}$$

We select the top-K posts with the highest depression risk scores as the input $X = \{p_1, ..., p_K\}$ for the user-level model.

3.2.2 Classification via Parameter Transfer

We extract and freeze the parameters of the well-optimized post-level post encoder M_{post} from the post-level task to preserve its learned semantic representations. Using this encoder, we encode the filtered user posts into embeddings $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_K] \in \mathbb{R}^{K \times d}$. Likewise, symptom descriptions are encoded into embeddings $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_{21}] \in \mathbb{R}^{21 \times d}$. We than adopt the symptom attention module from Liu et al. (2024), which employs projection matrices $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ to transform $\boldsymbol{H}$ into the key matrix $\boldsymbol{K} \in \mathbb{R}^{K \times d'}$ and the value matrix $\boldsymbol{V} \in \mathbb{R}^{K \times d'}$, and uses $\boldsymbol{W}_q$ to transform $\boldsymbol{Z}$ into the query matrix $\boldsymbol{Q} \in \mathbb{R}^{21 \times d'}$. The user embedding $\boldsymbol{U}$ is computed as follows:

$$\boldsymbol{A} = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d'}}) \tag{8}$$

$$\boldsymbol{U} = \boldsymbol{A}\boldsymbol{V} \tag{9}$$

where d' denotes the hidden dimension, and $\boldsymbol{A} \in \mathbb{R}^{21 \times K}$ represents the symptom-attention weight matrix. The user embedding $\boldsymbol{U}$ is subsequently mapped to $\boldsymbol{U}' \in \mathbb{R}^{21 \times d'}$ via a 12-layer transformer encoder. Finally, $\boldsymbol{U}'$ is pooled and fed into a two-layer classifier to produce the user label $\hat{y}_{\text{user}}$.

3.2.3 Comprehensive Interpretability

In addition to the user-level depression prediction result $\hat{y}_{\text{user}}$, the PAT framework can also output the following two types of explanatory information:

- **Post-Level Mental-State Trajectory:** Since we have already trained a effective post classifier in the post-level task, we leverage it to analyze each of the user's posts, thereby constructing a mental-state trajectory.
- **Key Symptoms:** From the symptom attention module, We extract $\boldsymbol{A} \in \mathbb{R}^{21 \times K}$ where each element $\boldsymbol{A}_{ij}$ reflects the relevance between the j-th post and the i-th BDI-II symptom. We select the top-3 symptoms with the highest attention weights as the user's symptom explanation.

4 Experimental Settings

4.1 Datasets

For the post-level dataset, we employ the depression severity four-class dataset constructed by Yang et al. (2021). The dataset contains 6,112 Chinese posts from Sina Weibo (a Chinese social media platform), each labeled with one of four depression levels: None, Slight, Moderate, Serious. The data are split into training, validation, and test sets in an 8:1:1 ratio.

For the user-level dataset, three datasets are chosen: SWDD (Cai et al., 2023), eRisk2017 (Losada and Crestani, 2016), and eRisk2018 (Losada et al., 2018). All three datasets are annotated with binary depression labels for users, based on user self-reports. The SWDD dataset is sourced from the Chinese social media platform Sina Weibo, while eRisk 2017 and eRisk 2018 are collected from the English Reddit forum. For each dataset, we split the data into training, validation, and test sets in an 8:1:1 ratio.

All the detailed dataset statistics are shown in Appendix A.3.

4.2 Implementation Details

Since our task comprises both Chinese and English datasets, we employ `ernie-3.0-base-zh` (ERNIE) (Sun et al., 2021) as the post encoder for PAT and all the competing methods. While its architecture aligns with that of BERT (Devlin et al., 2019), ERNIE supports both Chinese and English tasks. In the post-level task, we set $\tau = 0.05$, $w_{low} = 0,3$, $\alpha = 0.3$, and $\gamma = 4$. For the user-level task, we set $K = 16$ and $d' = 768$. The learning rate is fixed at 1e-5 for all experiments. The model is optimized with AdamW and trained for 10 epochs with a batch size of 16. The

Train	Method	Test: SWDD			Test: eRisk2017			Test: eRisk2018		
		F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
SWDD	HAN	84.87	82.56	87.32	35.29	60.00	25.00	33.33	63.50	22.72
	DeCapsNet	**90.01**	85.67	94.81	41.25	56.50	32.66	21.87	30.33	95.45
	Mental-BERT	85.86	79.26	93.66	——	——	——	——	——	——
	LLM	45.41	29.38	98.23	60.92	47.90	83.65	**58.72**	50.22	70.68
	LLM-COT	63.42	46.43	100.00	62.44	48.20	88.60	57.65	44.30	82.50
	PAT(BIG5)	**90.01**	89.79	90.39	**66.66**	55.55	83.33	**58.06**	45.00	81.81
	PAT(MBTI)	88.27	89.85	86.74	**72.00**	69.23	75.00	55.81	57.14	54.54
eRisk 2017	HAN	7.52	56.00	4.03	64.00	61.53	66.66	62.74	55.17	72.73
	DeCapsNet	35.37	30.25	42.57	64.28	56.25	75.00	62.22	60.86	63.63
	Mental-BERT	——	——	——	62.06	52.94	75.00	**70.37**	59.37	86.36
	LLM	46.89	31.28	93.66	58.32	42.85	92.30	68.09	64.00	77.73
	LLM-COT	64.68	52.31	84.72	48.49	40.00	61.54	65.33	60.87	70.50
	PAT(BIG5)	**73.06**	87.85	62.53	**67.80**	69.40	66.67	65.38	56.66	77.27
	PAT(MBTI)	**69.43**	93.20	55.33	**66.67**	60.00	75.00	**69.39**	62.96	77.27
eRisk 2018	HAN	10.36	51.28	5.76	——	——	——	61.12	55.86	68.18
	DeCapsNet	20.18	38.61	13.66	——	——	——	66.67	68.18	65.22
	Mental-BERT	——	——	——	——	——	——	68.09	64.00	72.73
	LLM	51.71	35.32	96.45	——	——	——	65.31	59.26	72.73
	LLM-COT	**59.41**	46.43	82.46	——	——	——	51.43	37.50	81.82
	PAT(BIG5)	54.47	97.76	37.75	——	——	——	**72.37**	68.11	77.27
	PAT(MBTI)	**65.81**	90.86	51.59	——	——	——	**69.57**	66.67	72.73

Table 1: Main results of experiments in the user-level depression detection. The experiments include both within-dataset and cross-dataset tests. The F1 of the top-2 model in each group are highlighted in bold. The eRisk2018 to eRisk2017 experiment was omitted to avoid data leakage, given that eRisk2018 encompasses the entire eRisk2017 dataset.

checkpoint achieving the best performance on the validation set is selected for final testing. To ensure robustness, each model is trained with three different random seeds. All experiments are conducted on an NVIDIA GTX 3090 GPU (24GB).

4.3 Competing Methods

We compare our proposed method in the user-level depression detection task with the following ones:

- **HAN** (Zhang et al., 2022): This method employs post screening followed by a hierarchical attention network for classification.
- **DeCapsNet** (Liu et al., 2024): Similar to **HAN**, this approach replaces the user-level hierarchical module with a capsule network and additionally incorporates contrastive learning loss for optimization.
- **Mental-BERT** (Ji et al., 2022): This method pre-trains BERT on a psychology-specific corpus to improve domain adaptation. We replace our post encoder with its parameters to compare the effectiveness of our encoder optimization. For the Chinese dataset SWDD, we use **Chinese Mental-BERT** (Zhai et al., 2024).
- **LLM**: We perform depression detection via prompting a large language model, providing the BDI-II diagnostic criteria and 2-shot examples from the train set. Concretely, we adopt `Qwen-plus-2025-07-28` (Yang et al., 2025) for testing. In addition, we also test the version that uses the Chain-of-Thought (COT) prompt (Wei et al., 2022).

5 Results

In the experiments , we focus on the five research questions (RQs) below:

- **RQ1:** How dose the proposed PAT framework perform compared to existing baselines?
- **RQ2:** Can post-level training improve user-level performance?

- **RQ3:** How effective are the personality-driven data augmentation and fine-grained contrastive learning? And why?
- **RQ4:** How accurate and clinically meaningful are the explanations generated by PAT?

5.1 Main Results

To answer **RQ1**, we evaluate user-level depression detection under both within-dataset and cross-dataset settings. The within-dataset setting involves training and testing on the same dataset, while the cross-dataset setting assesses models' generalization ability by training on one dataset and testing on another. The main results are summarized in Table 1.

Within-Dataset Performance: The proposed PAT framework (with either BIG5 or MBTI theory) achieves the best performance across all three datasets. On the SWDD dataset, its performance is on par with that of DeCapsNet. On the English datasets (eRisk 2017 and 2018), PAT yield statistically significant improvements over other strong baselines. Additionally, the BIG5-based implementation consistently exceeds the MBTI-based counterpart in all within-dataset evaluations.

Cross-Dataset Performance: In cross-dataset evaluations, PAT demonstrates superior generalization, outperforming most baselines in the majority of the dataset pairs. Performance is marginally lower than the top baseline only in the SWDD to eRisk2018 and eRisk2017 to eRisk2018 experiments. Notably, in the SWDD to eRisk2017 cross-dataset experiment, all methods are trained only on Chinese data. While the performance of HAN and DeCapsNet deteriorate substantially on the English (eRisk2017) test set, the PAT(MBTI) achieves the best performance, even surpassing its own within-dataset results. Our further analysis indicates that the post encoder is capable of capturing depression patterns in English text, even without explicit English training. A detailed discussion is provided in Appendix A.2.

5.2 Ablation Study

We perform an ablation study for the user-level task with four settings: (1) **w/o T**: Removing the transfer learning module, using only the original pre-trained model parameters. (2) **w/o FGCL**: Replacing the fine-grained contrastive learning module with a standard supervised contrastive learning approach. (3) **w/o PA**: Removing the data augmentation. (4) **w/o Freeze**: Unfreezing the post encoder for fine-tuning. The results are presented in Table 2.

For **RQ2**, the **w/o T** setup results in a substantial performance decline, which confirms that the post-level auxiliary task is crucial. The corresponding drop in the **w/o Freeze** setting indicates that fine-tuning the encoder with the coarse supervision signal from the user-level task is suboptimal and risks introducing noise, as explained in Section 1.

Method	SWDD F1	eRisk2017 F1	eRisk2018 F1
PAT	90.01	67.80	72.37
w/o T	-11.21	-15.49	-9.38
w/o FGCL	-1.32	-1.84	-0.42
w/o PA	-2.34	-4.89	-2.05
w/o Freeze	-1.20	-2.72	-2.65

Table 2: Results of the ablation analysis on the user-level task.

For **RQ3**, The performance drop in the user-level ablation study, observed when either fine-grained contrastive learning or personality-driven augmentation is removed, confirms the effectiveness of both components. To investigate this more deeply, we conducted an additional ablation study on the post-level task. The results, presented in Table 3, yield consistent findings, further validating the utility of both modules.

Furthermore, we visualize the embedding space to compare the embedding space of different contrastive learning methods. As shown in Figure 4, although supervised contrastive learning can achieve clear separation among the four class clusters, the distances from the "None", "Slight", and "Moderate" clusters to the "Serious" cluster are almost identical. By comparison, in the embedding space learned by fine-grained contrastive learning, the distribution of representations exhibits a clear pattern: the clusters from right to left in the figure correspond to progressively increasing levels of depression severity. This indicates that fine-grained contrastive learning can more effectively optimize

Method	Acc	F1
our	75.12	73.84
w/o FGCL	-1.93	-1.11
w/o PA	-2.63	-2.58

Table 3: Results of the ablation analysis on the post-level task.

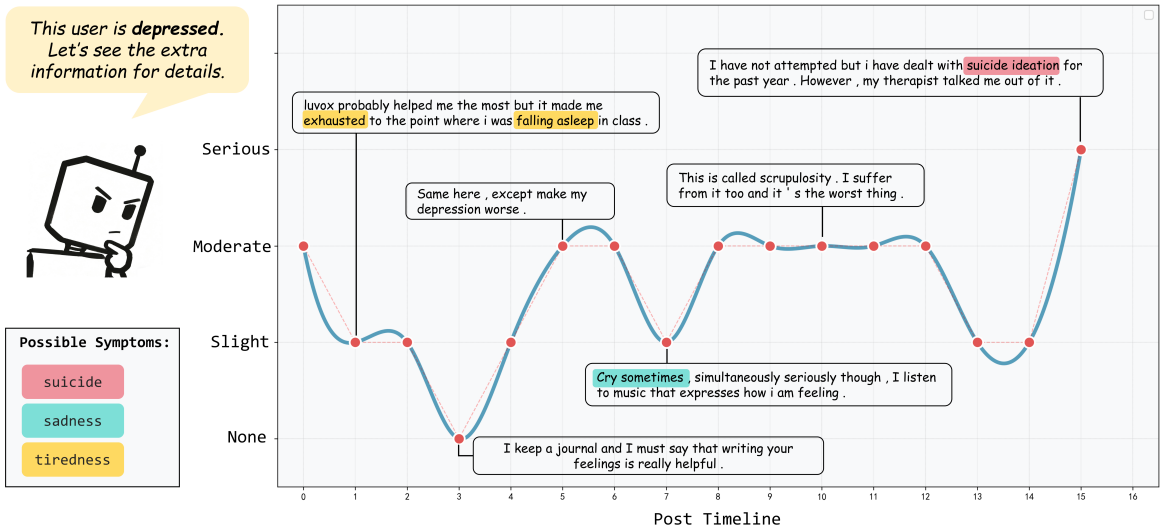
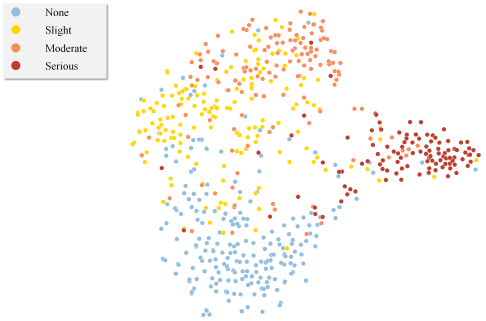
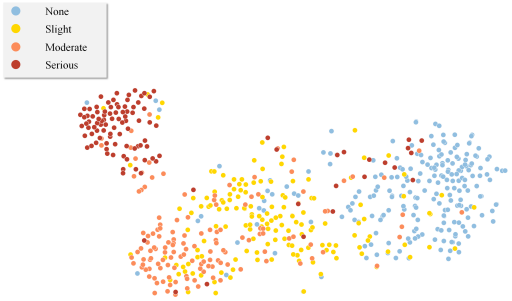


Figure 3: A case study demonstrating the comprehensive interpretability of the PAT framework.

the embedding space, thereby enhancing the ability of the post encoder.



(a) Trained with Supervised Contrastive Learning



(b) Trained with Fine-Grained Contrastive Learning

Figure 4: Visualization of the embedding space of different contrastive learning methods.

5.3 Case Study

For **RQ4**, a detailed user case analysis is provided in Figure 3. The PAT framework not only flags this user as depressed but also outputs the longitudinal mental-state trajectory and key symptoms. The severity trajectory shows the user's depression level long oscillated between "Slight" and "Moderate", before escalating to "Serious" in the latest post. The post content validates this trend against the defined criteria: "Slight" (mild, short-term symptoms); "Moderate" (severe, long-term symptoms); "Serious" (suicidal ideation) (Yang et al., 2021). The key symptoms (suicide, sadness, tiredness) also match the content. All the detailed information can support experts in diagnosis and intervention. Additional cases are provided in Appendix A.5.

6 Conclusion

In this paper, we introduce PAT, a personality-driven augmentation and transfer learning framework, designed to overcome the bottleneck in the post encoder for depression detection. PAT introduces an auxiliary post-level task that effectively trains the post encoder using personality-driven augmentation and fine-grained contrastive learning. Subsequently, this well-optimized encoder is transferred to the user-level model. These techniques enable the post encoder to better capture depression semantics while remaining robust against coarse user-level signals. Extensive experiments validate PAT's effectiveness and strong generalization capability across datasets. Moreover, by providing detailed evidence, PAT enhances the trustworthiness of its predictions and offers actionable insights for diagnosis and intervention.

7 Limitations

While the proposed PAT framework demonstrates notable advantages in both performance and generalization capability, this study still has certain limitations, which we hope will provide direction for future work.

Limitations in Data Resources: The post encoder of PAT was optimized on Chinese data, whereas the user-level evaluation spanned both Chinese (SWDD) and English (eRisk series) datasets. While training with English post-level data could lead to further performance gains and more robust validation, the current lack of annotated English post-level depression datasets remains a primary constraint.

Lack of Expert Validation for Interpretability: Although PAT provides intuitive interpretability evidence, and case studies show consistency with the content of the posts, this evidence has not yet been systematically evaluated by domain experts. Future work should involve large-scale human verification, particularly incorporating assessments by psychology experts, to rigorously quantify its clinical validity and practical utility.

Breadth of Application for Psychological Methods: This study validated the use of two personality theories (BIG5 and MBTI) for data augmentation. An important direction for extension is to apply such psychology-driven augmentation strategies to a broader range of mental health computing tasks and to explore more diverse theoretical frameworks.

8 Ethical Statement

The goal of this study on detecting depression using social media data is to help individuals with depression recover as early as possible and prevent severe consequences such as suicidal behavior. However, we also recognize that this research involves certain ethical risks. To mitigate these risks, we issue the following statement:

Data Source and Privacy Protection: This study exclusively utilizes publicly available datasets, and their use was formally approved by the publishers following the submission of a usage application form. Through a combination of automated tools and manual review, we have rigorously de-identified all data, permanently removing all personally identifiable information such as usernames, nicknames, geographical locations, and contact details. Any sample texts cited in the paper have been paraphrased to prevent any possibility of tracing them back to specific individuals.

Nature of the Model and Usage Restrictions: The model developed in this study is intended solely as a preliminary detection tool for research purposes and is in no way a professional diagnostic tool. It cannot and should not replace formal assessment, diagnosis, or treatment by qualified mental health professionals. The output of the model is limited to academic research purposes and is strictly prohibited from being used for any form of individual assessment or in contexts that may lead to discrimination.

Awareness of Risks and Responsibility: We are fully aware that algorithms may contain biases and inaccuracies, and that misunderstandings and discrimination regarding mental health issues exist in society. We take full responsibility for the security of the data to prevent any breaches that could cause further harm to potentially affected individuals.

References

Navneet Agarwal, Kirill Milintsevich, Lucie Metivier, Maud Rotharmel, Gaël Dias, and Sonia Dollfus. 2024. Analyzing symptom-based depression level estimation through the prism of psychiatric expertise. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 974–983, Torino, Italia. ELRA and ICCL.

Usman Ahmed, Gautam Srivastava, Unil Yun, and Jerry Chun-Wei Lin. 2022. Eandc: An explainable attention network based deep adaptive clustering model for mental health treatment. *Future Generation Computer Systems*, 130:106–113.

Aaron T Beck. 1996. Manual for the beck depression inventory-ii. *(No Title)*.

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

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

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.

Yan Ding, Xuemei Chen, Qiming Fu, and Shan Zhong. 2020. A depression recognition method for college students using deep integrated support vector algorithm. *IEEE access*, 8:75616–75629.

Prasadith Kirinde Gamaarachchige and Diana Inkpen. 2019. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019)*, pages 54–64.

Md Tazmim Hossain, Md Arafat Rahman Talukder, and Nusrat Jahan. 2021. Social networking sites data analysis using nlp and ml to predict depression. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.

Carl Jung and John Beebe. 2016. *Psychological types*. Routledge.

Harnain Kour and Manoj K Gupta. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bidirectional lstm. *Multimedia Tools and Applications*, 81(17):23649–23685.

Antonis Koutsoumpis, Janneke K Oostrom, Djurre Holtrop, Ward Van Breda, Sina Ghassemi, and Reinout E de Vries. 2022. The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the big five and the linguistic inquiry and word count (liwc). *Psychological Bulletin*, 148(11-12):843.

Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression detection on social media with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2155–2171, Suzhou (China).

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

Han Liu, Changya Li, Xiaotong Zhang, Feng Zhang, Wei Wang, Fenglong Ma, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2024. Depression detection via capsule networks with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22231–22239.

David E Losada and Fabio Crestani. 2016. 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.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk: early risk prediction on the internet. In *International conference of the cross-language evaluation forum for european languages*, pages 343–361. Springer.

S Ram Narayanan, S Babu, and Amal Thandayantavida. 2022. Detection of depression from social media using deep learning approach. *Journal of positive school psychology*, 6(4):4909–4915.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2023. Semantic similarity models for depression severity estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16104–16118, Singapore.

Federico Ravenda, Seyed Ali Bahrainian, Andrea Raballo, Antonietta Mira, and Noriko Kando. 2025. Are LLMs effective psychological assessors? leveraging adaptive RAG for interpretable mental health screening through psychometric practice. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8975–8991, Vienna, Austria.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, and 1 others. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Waleed Bin Tahir, Shah Khalid, Sulaiman Almutairi, Mohammed Abohashrh, Sufyan Ali Memon, and Jawad Khan. 2025. Depression detection in social media: A comprehensive review of machine learning and deep learning techniques. *IEEE Access*.

Zannatun Nayem Vasha, Bidyut Sharma, Israt Jahan Esha, Jabir Al Nahian, and Johora Akter Polin. 2023. Depression detection in social media comments data using machine learning algorithms. *Bulletin of Electrical Engineering and Informatics*, 12(2):987–996.

Bichen Wang, Yuzhe Zi, Yixin Sun, Hao Yang, Yanyan Zhao, and Bing Qin. 2025. End-to-end learnable psychiatric scale guided risky post screening for depression detection on social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4054–4066, Suzhou, China.

Bichen Wang, Yuzhe Zi, Yanyan Zhao, Pengfei Deng, and Bing Qin. 2024. ESDM: Early Sensing depression model in social media streams. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6288–6298, Torino, Italia. ELRA and ICCL.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiageng Wu, Xian Wu, Yining Hua, Shixu Lin, Yefeng Zheng, and Jie Yang. 2023. Exploring social media for early detection of depression in covid-19 patients. In *Proceedings of the ACM Web Conference 2023*, pages 3968–3977.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tingting Yang, Fei Li, Donghong Ji, Xiaohui Liang, Tian Xie, Shuwan Tian, Bobo Li, and Peitong Liang. 2021. Fine-grained depression analysis based on chinese micro-blog reviews. *Information Processing & Management*, 58(6):102681.

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

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Yang, and Guanghui Fu. 2024. Chinese mentalbert: Domain-adaptive pretraining on social media for chinese mental health text analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10574–10585.

Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. 2025. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9927–9944, Vienna, Austria.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2022. Psychiatric scale guided risky post screening for early detection of depression. *arXiv preprint arXiv:2205.09497*.

Tong Zheng, Yanrong Guo, and Richang Hong. 2024. Cascade large language model via in-context learning for depression detection on chinese social media. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 353–366. Springer.

A Appendix

A.1 Prompt Templates

The prompt template adopted for the personality-driven augmentation is shown in Figure 5.

A.2 Analysis of Language Transfer

To investigate why the PAT framework, despite not being trained on English, still achieves impressive performance on English datasets (as demonstrated in the SWDD to eRisk2017 experiment), we conducted a visualization study. Specifically, we used the post encoder to encode both sentences from the eRisk2017 test set and posts from the post-level dataset, followed by dimensionality reduction and visualization using t-SNE, as shown in Figure 6.

It can be observed that although the posts in the eRisk2017 dataset are not perfectly aligned with the sample points from the Chinese post-level dataset, the "Depressed" data points overall exhibit higher density near the "Moderate" and "Serious" clusters, and lower density near the "None" and "Slight" clusters. This indicates that the post encoder, despite being trained without English data, can still capture the depression distribution pattern to some extent, thereby demonstrating strong cross-lingual transfer capability. It should be noted that although some "Depressed" points also appear near the "None" cluster, we labeled all posts from depressed users in the user-level dataset as "Depressed", which inevitably introduces false-positive samples. This is exactly the issue we highlighted in Figure 1.

A.3 Dataset Statistics

The statistics of the post-level and user-level datasets are shown in Table 4 and Table 5, respectively. Specially, the post-level dataset originally comprises five classes: None, Slight, Moderate, Severe, Very Severe. Considering label balance and the fact that most medical depression scales adopt four-level classifications, we merge the “Severe” and “Very Severe” labels into a single “Serious” category.

A.4 Symptom Descriptions

To screen for depressive posts, we converted the descriptions of the 21 questions from the Beck

You are a psychology expert and writing style conversion assistant. Your task is to rewrite the input blog post sentence based on the user input and the Big Five personality traits, assuming you are the speaker, to make it conform to the typical expression style of that personality, without changing the original sentence's level of depressive symptoms. Depressive symptoms refer to PHQ-9, including loss of interest, low mood, sleep disturbances, appetite changes, low self-esteem, concentration difficulties, and self-harm.

Rules:
Depressive tendency includes four levels: asymptomatic; mild and transient symptoms; severe and chronic symptoms; suicidal ideation or behavior. Do not alter the original sentence's tendency.

Big Five personality style adjustments (10 types):
1. ***High extraversion***: *Use more social, active expressions (e.g., "I want to talk to a friend").*
2. ***Low extraversion***: *More introverted, concise (e.g., "I need some time alone").*
3. ***High agreeableness***: *Gentle, empathetic (e.g., "I understand how you feel").*
4. ***Low agreeableness***: *Direct, critical (e.g., "The problem is obvious").*
5. ***High conscientiousness***: *Structured, responsibility-oriented (e.g., "I should make a plan").*
6. ***Low conscientiousness***: *Casual, unplanned (e.g., "Whatever, doesn't matter").*
7. ***High neuroticism***: *Emotional, anxious (e.g., "I always feel like something will go wrong").*
8. ***Low neuroticism***: *Calm, stable (e.g., "Things will work out eventually").*
9. ***High openness***: *Abstract, metaphorical (e.g., "Life is like a fog").*
10. ***Low openness***: *Concrete, direct (e.g., "Work hasn't been going well lately").*

Output requirements:
- Return only the rewritten sentences, no explanations or additional notes. Output for all ten cases, one per line. For example, the first line should be "High extraversion: [Rewrite for High Extraversion]"
- Ensure the rewritten sentences are natural, fluent, and conform to everyday language habits.

User Input:
[Post Text]

You are a psychology expert and a writing style conversion assistant. Your task is to rewrite the input blog post sentence into a natural expression on social media, imitating the daily posting style of real users on platforms like Moments and Weibo, based on the user input and the eight cognitive function dimensions of MBTI, while strictly maintaining the original sentence's level of depressive symptoms.

Rules:
Maintain the Four Levels of Depressive Tendency: Strictly preserve the depressive severity of the original sentence (asymptomatic / mild and transient / severe and chronic / suicidal ideation)

MBTI personality style adjustments (8 types)
1. ***Extroverted (E)***: *Socialized expression, with a tendency toward descriptions related to interaction with others, e.g., "Refused all invites from friends, really not in the mood to socialize."*
2. ***Introverted (I)***: *Introspective expression, focusing on internal feelings, e.g., "When I'm alone, this sense of helplessness feels even more obvious."*
3. ***Sensing (S)***: *Concrete, practical description, focusing on details and reality, e.g., "Haven't slept for three nights, dark circles are so bad I can't face people."*
4. ***Intuition (N)***: *Abstract, associative expression, focusing on possibilities and meaning, e.g., "Feels like the future is all fog, can't find the way."*
5. ***Thinking (T)***: *Logical-analytical expression, describing problems objectively, e.g., "Analyzed it a bit, this state is already affecting my normal work."*
6. ***Feeling (F)***: *Value- and care-oriented expression, emphasizing feelings and empathy, e.g., "I know this isn't good, but just can't control my emotions."*
7. ***Judging (J)***: *Plan-oriented expression, emphasizing order and decisiveness, e.g., "All my original plans are messed up, now I can't get anything done."*
8. ***Perceiving (P)***: *Flexible, adaptive expression, maintaining openness and casualness, e.g., "Thought of just going with the flow, but things seem to be getting worse."*

Output requirements:
- Return only the rewritten sentences, no explanations or additional notes. Output for all ten cases, one per line. For example, the first line should be "Extroverted: [Rewrite for Extroverted]"
- Ensure the rewritten sentences are natural, fluent, and conform to everyday language habits.

User Input:
[Post Text]

Figure 5: Personality-Driven Augmentation Prompt Templates. The templates are built on two personality theories: the Big Five (ten traits from high/low tendencies of its five dimensions) and MBTI (eight Jungian cognitive functions). Each template guides the LLM to rewrite a sentence in the linguistic style of a specified personality trait while preserving its original depressive-symptom description, with a one-shot example provided per trait.

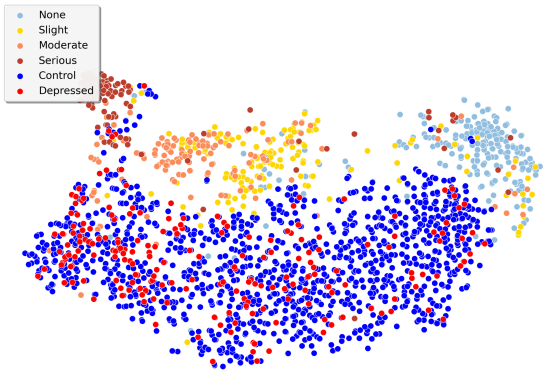


Figure 6: Visualization of the embedding space of the post-level dataset and eRisk2017 dataset. The data points labeled "None," "Slight," "Moderate," and "Serious" are sourced from the post-level dataset. The "Depressed" and "Control" data points are taken from the eRisk2017 dataset, where all posts from depressed users are labeled as "Depressed," and all posts from control users are correspondingly labeled as "Control."

	Train	Test	Val
No. of posts(None)	1922	218	210
No. of posts(Slight)	1193	151	158
No. of posts(Moderate)	1107	135	144
No. of posts(Serious)	670	107	100
Sum	4892	611	612

Table 4: Label Distribution and Data Split of the Post-Level Depression Severity Classification Dataset.

Depression Inventory-II (BDI-II) into first-person symptom descriptions. All symptom descriptions are given in Table 6.

A.5 More Cases

A Case of Healthy User Figure 7 presents the interpretability analysis results of PAT for a normal user. It can be observed that none of the user's posts reach the "Serious" level, with the majority falling below "Slight," indicating a relatively low risk of depression. In the symptom analysis, the user indeed exhibits linguistic patterns associated with agitation and pessimism. However, suicide is not mentioned in the user's posts. The high attention weight given to this category may be attributed to the fact that suicide is a critical factor in depression diagnosis.

Cases of Model Misidentification In this section, we present several error cases for analysis. As shown in Figure 8, a depressed user was incorrectly classified as normal. The depression trajectory of this user's posts indicates that their depression level remained at "Moderate" for an extended period but never reached the "Serious" level, which likely contributed to the missed detection. The results of the symptom attribution analysis are largely consistent with the user's expressions, except for the absence of any mention of "suicide."

Another case in which a normal user was mis-

	SWDD			eRisk 2017			eRisk 2018		
	Train	Test	Val	Train	Test	Val	Train	Test	Val
No. of Users (Dep)	2,968	347	396	110	12	13	165	27	22
No. of Users (Con)	15,604	1,976	1,925	599	77	76	1,199	144	149
No. of Total Users	18,572	2,323	2,321	709	89	89	1,364	171	171

Table 5: Label Distribution and Data Split of the User-Level Datasets.

	Symptom	Descriptions
1	Sadness	I feel so sad or unhappy
2	Pessimism	I feel the future is hopeless
3	Past Failure	I feel I am a complete failure as a person
4	Loss of Pleasure	I am dissatisfied or bored with everything
5	Guilty Feelings	I feel guilty all of the time
6	Punishment Feelings	I feel I am being punished
7	Self-Dislikes	I hate myself
8	Self-Criticalness	I blame myself all the time for my faults
9	Suicide	I would like to kill myself
10	Crying	I cry all the time now
11	Agitation	I feel irritated all the time
12	Loss of Interest	I have lost most of my interest in other people
13	Indecisiveness	I have greater difficulty in making decisions
14	Worthlessness	I believe that I am worthless
15	Loss of Energy	I have to push myself very hard to do anything
16	Sleep Question	I can't sleep well
17	Irritability	I get angry very easily these days.
18	Changes in Appetite	My appetite has changed drastically
19	Concentration Difficulty	I can't concentrate.
20	Tiredness	I feel tired all the time
21	Loss of Interest in Sex	I am much less interested in sex now

Table 6: Symptom descriptions adopted for screening and the symptom-attention module.

classified as depressed is presented in Figure 9. Although the user's depression trajectory shows prolonged "Moderate" levels and occasional spikes to "Serious", further analysis reveals that many of the posts, especially those reaching "Serious", actually contain quotes expressing negative emotions from comics or movies, which likely led to the model's misjudgment. However, it is worth noting that some of the user's posts also reflect experiences of domestic violence and prolonged low mood, indicating a state of emotional distress that carries a risk of further deterioration. Therefore, while the classification may be technically inaccurate in a diagnostic sense, the model's identification of this case could still be considered meaningful from a risk-aware perspective.

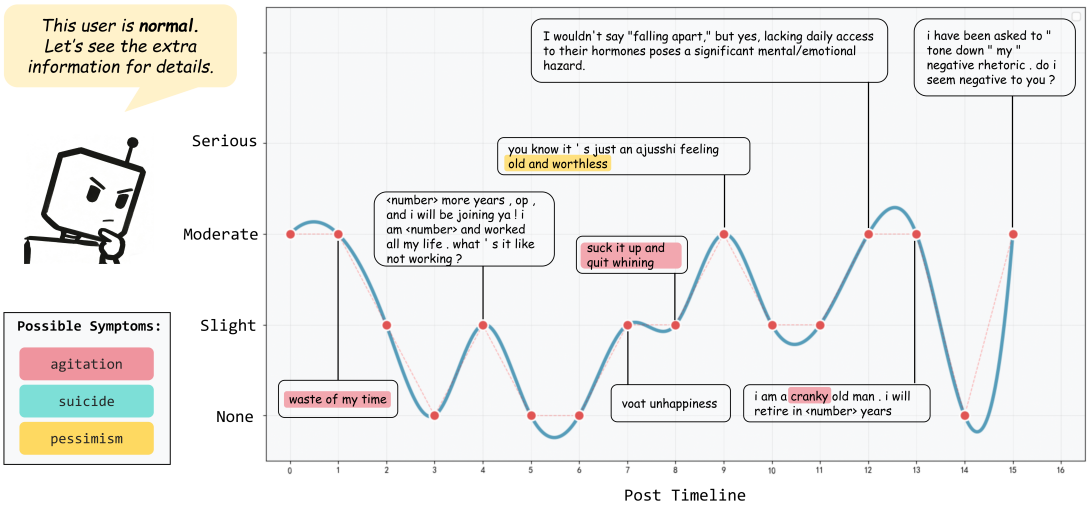


Figure 7: A case of a normal user.

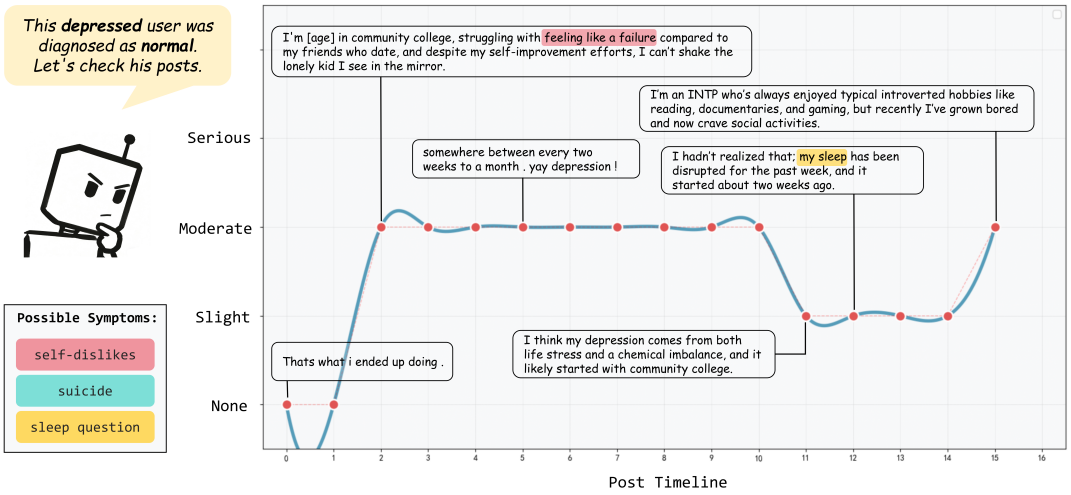


Figure 8: A case of a depressed user that was missed in detection.

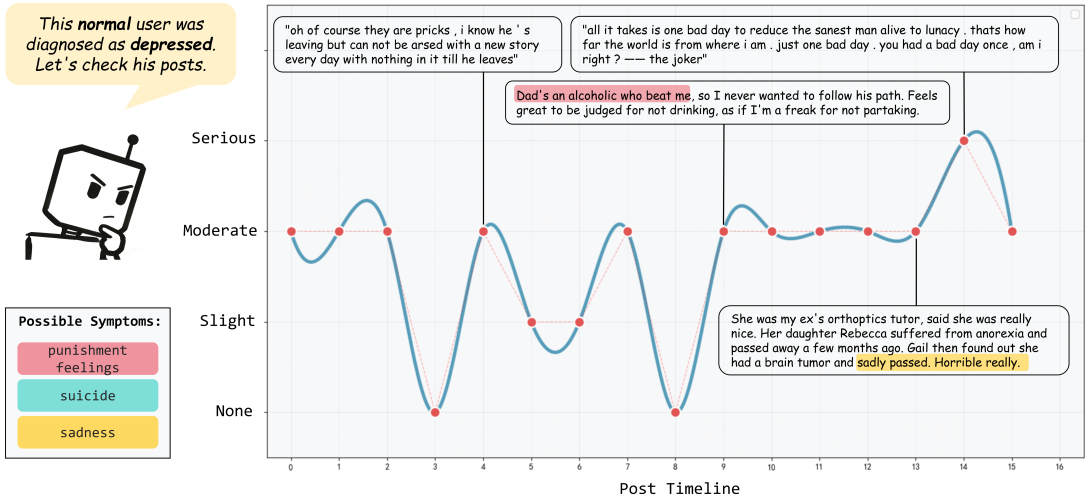


Figure 9: A case of a normal user misdiagnosed with depression.