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 that affects millions worldwide, with early detection crucial for effective intervention. While existing methods have achieved promising results in automated depression detection, they face two key limitations: (1) noise in post screening: symptom-based semantic screening is easily misled by superficial lexical patterns, thereby introducing false-positive posts; (2) weak optimization of the post encoder: relying solely on coarse user-level supervision signals prevents models from capturing depressive cues within individual posts. To address these limitations, we propose PAT, a **P**ersonality-driven **A**ugmentation and **T**ransfer learning framework. In the first phase, PAT optimizes a post-level encoder via a fine-grained contrastive learning objective, utilizing LLM-based data augmentation driven by classic personality theories to alleviate data scarcity. In the second phase, the trained post-level scorer is deployed for high-quality post filtering, while its encoder parameters are transferred for user-level tasks. Extensive experiments demonstrate that PAT consistently outperforms existing methods. Furthermore, PAT delivers multi-granular interpretability by tracing post-level mental-state trajectories and highlighting key symptoms, thereby offering sufficient diagnostic evidence for our detection results.

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 people worldwide 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.

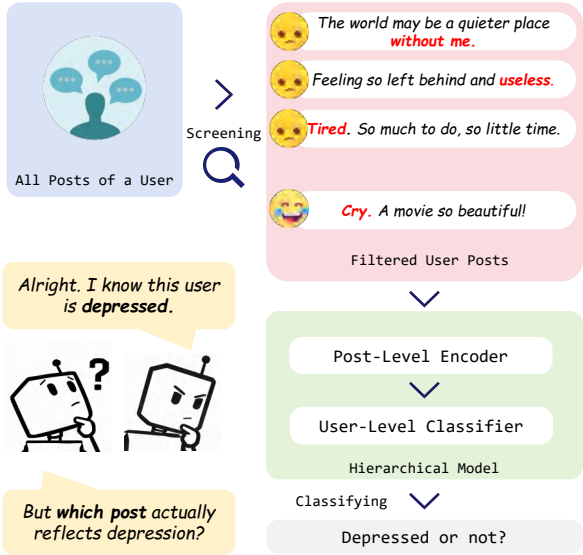


Figure 1: A typical Screen-and-Classify framework for social media user-level depression detection. The hierarchical model is trained solely on binary supervision (depressed or non-depressed), which prevents the post encoder from capturing fine-grained depressive cues in each post and thus limits its 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 feelings online, providing extensive data to develop automated tools for depression detection (Ding et al., 2020; Wang et al., 2024; Tahir et al., 2025). 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 frame-

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

work comprises two steps: (1) Post Screening: retrieving posts related to depression using symptom descriptions as queries; and (2) Hierarchical Classification: encoding the posts and aggregating embeddings for classification.

However, both steps in this framework face bottlenecks: First, post screening may introduce false positives that contain symptom keywords but express positive meanings. Second, the optimization relies solely on sparse user-level supervision, providing insufficient guidance for the post encoder. 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 using Large Language Models (LLMs) for post-level information extraction and explanation generation (Lan et al., 2025; Zheng et al., 2024). Although 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 a two-phase **P**ersonality-driven **A**ugmentation and **T**ransfer learning framework (PAT), whose overall structure is shown in Figure 2. In the first phase, PAT trains a post-level scorer to evaluate depressive severity, which is subsequently leveraged to perform score-based filtering on user-level inputs while its encoder component is transferred for user-level training. To alleviate the inherent scarcity of post-level data, we draw inspiration from psycholinguistic findings that personality drives distinct linguistic manifestations (Koutsoumpis et al., 2022). Specifically, we employ LLMs for personality-driven data augmentation based on personality theories, including the Big Five (Costa and McCrae, 2008) and MBTI (Jung and Beebe, 2016), thus expanding the diversity of data in various expression styles. Furthermore, a fine-grained contrastive learning objective is designed to effectively utilize both the augmented and labeled samples for robust representation learning. Beyond performance enhancements, PAT distinguishes itself by providing multi-granular interpretability: it delivers user-level predictions, traces post-level mental-state trajectories, and highlights key symptoms, thus offering comprehensive diagnostic evidence. In summary, the contributions of this paper are threefold:

- We propose a novel transfer learning framework, PAT. Extensive experiments on multiple benchmarks demonstrate that PAT outperforms existing strong baselines and can serve as a plug-and-play component to enhance other methods.
- We design a personality-driven data augmentation method with a corresponding fine-grained contrastive learning algorithm. Experiments demonstrate that these methods effectively optimize the embedding space of the post encoder, while revealing the limitations of conventional supervised contrastive learning.
- 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 improve 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; 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 classification, 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 end-to-end 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.

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

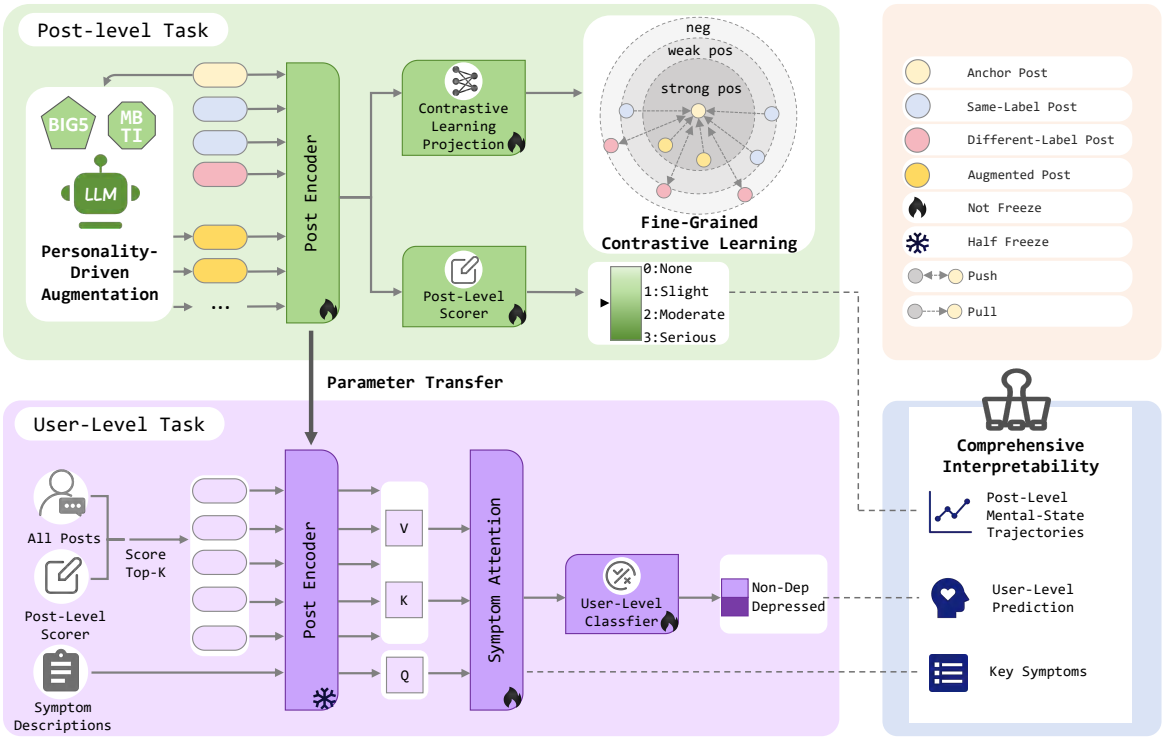


Figure 2: An overview of the PAT framework: it first performs post-level optimization and then transfers the post scorer's encoder to the user-level task for screening and encoding. For post-level optimization, PAT introduces **Personality-Driven Augmentation** and novel **Fine-grained Contrastive Learning**. 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.

non-parametric. 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 improve 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) and Zhai et al. (2024) pre-trained a domain-specialized Mental-BERT in the psychological corpora. Wu et al. (2023) transferred knowledge from sentiment classification to depression detection via distillation. 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 using 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 a 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 depression-level scorer trained in the post-level task is deployed for user-level post screening, while its encoder is 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 formulate a regression task that estimates a depression-intensity score $r \in [0, 3]$, where a higher value indicates increased severity at four progressive levels: None, Slight, Moderate, and Serious.

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. 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 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_i into a text embedding. Then, we utilize a two-layer linear regression head to produce an unnormalized score $\tilde{r}_i$. The Mean Squared Error (MSE) loss is computed as:

$$\hat{r}_i = 3 \cdot \text{Sigmoid}(\tilde{r}_i) \tag{1}$$

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(r_i - \hat{r}_i)^2 \tag{2}$$

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:

$$\mathcal{L}_{cl} = \frac{1}{B}\sum_{i\in I}\frac{-1}{|\Gamma(i)|}\sum_{j\in\Gamma(i)}\log\frac{w_{ij}\cdot e^{\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau}}{\sum_{k\in I\setminus i}e^{\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k)/\tau}} \tag{3}$$

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, $\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)$ 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:

$$d_{ij} = 1 - \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j) \tag{4}$$

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

$$\overline{d}_i = \frac{1}{|\tilde{Z}|}\sum_{\tilde{\boldsymbol{z}}\in\tilde{Z}}(1 - \text{sim}(\boldsymbol{z}_i, \tilde{\boldsymbol{z}})) \tag{5}$$

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

$$w_{ij} = \begin{cases} w_{low} + (1 - w_{low})e^{-\gamma\frac{d_{ij}-\overline{d}_i}{d_{ij}}} & \text{if } d_{ij} > \overline{d}_i \\ 1 & \text{if } d_{ij} \leq \overline{d}_i \end{cases} \tag{6}$$

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:

$$\mathcal{L}_{total} = \mathcal{L}_{mse} + \alpha\mathcal{L}_{cl} \tag{7}$$

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 $P = \{p_1, \ldots, p_n\}$, previous methods filter relevant posts based on semantic similarity to symptom descriptions $\boldsymbol{S} = \{\boldsymbol{s_1}, ..., \boldsymbol{s_m}\}$ from the Beck Depression Inventory-II (BDI-II) (Beck, 1996). In contrast to relying on semantic matching, we directly adopt the trained scorer on the post-level task to assign a depression-level score $\hat{r}_i$ to each post. We select the top-K posts with the highest 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} except for the last two layers 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 $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_m] \in \mathbb{R}^{m \times d}$. We then adopt the symptom attention module inspired by 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 $\boldsymbol{Q} \in \mathbb{R}^{m \times d'}$. The user embedding $\boldsymbol{U}$ is computed as follows:

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

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

where d' denotes the hidden dimension, and $\boldsymbol{A} \in \mathbb{R}^{m \times K}$ represents the symptom-attention weight matrix. The user embedding $\boldsymbol{U}$ is subsequently mapped to $\boldsymbol{U}' \in \mathbb{R}^{m \times d'}$ via a 6-layer transformer encoder. Finally, $\boldsymbol{U}'$ is mean-pooled and fed into a two-layer linear 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 an effective depression-level scorer 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 $\boldsymbol{A} \in \mathbb{R}^{m \times K}$ 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). The dataset is split into training, validation, and test sets in a 8:1:1 ratio. For personality-driven data augmentation, we use `Qwen3.5-Plus` to generate the augmented samples. Due to the differing augmentation scales of the two methods (10× for Big Five vs. 8× for MBTI), we randomly retain eight augmented posts per instance for the Big Five to ensure a balanced comparison.

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 their self-reports. The SWDD dataset is sourced from the Chinese social media platform Sina Weibo, while eRisk2017 and eRisk2018 are collected from the English Reddit forum. For each dataset, we split the data into training, validation, and test sets in a 3:1:1 ratio.

All detailed data statistics and augmented examples are shown in Appendix A.2.

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 $\tau = 0.5$, $w_{\text{low}} = 0.3$, $\alpha = 0.5$, 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. Models are optimized with AdamW and trained for 10 epochs with a batch size of 16. The checkpoint achieving the best performance on the validation

Methods	Dataset	F1	Precision	Recall	AUC
HAN	eRisk2017	0.570 ± 0.013	0.564 ± 0.030	0.575 ± 0.036	0.879 ± 0.002
DeCapsNet		0.579 ± 0.013	0.563 ± 0.038	0.586 ± 0.024	0.891 ± 0.010
E2LPS		0.575 ± 0.009	0.559 ± 0.018	0.594 ± 0.040	0.884 ± 0.010
Mental-Bert		0.629 ± 0.012	0.595 ± 0.034	0.667 ± 0.018	0.891 ± 0.002
GPT-5.4		0.632 ± 0.009	0.702 ± 0.016	0.575 ± 0.028	-
Qwen3.5-Plus		0.527 ± 0.014	0.389 ± 0.024	0.815 ± 0.036	-
PAT (BIG5)		**0.640** ± 0.012	0.607 ± 0.023	0.680 ± 0.024	**0.902** ± 0.002
PAT (MBTI)		0.637 ± 0.015	0.610 ± 0.020	0.667 ± 0.038	0.898 ± 0.004
HAN	eRisk2018	0.610 ± 0.013	0.623 ± 0.017	0.600 ± 0.026	0.896 ± 0.003
DeCapsNet		0.631 ± 0.015	0.615 ± 0.041	0.649 ± 0.033	0.904 ± 0.007
E2LPS		0.566 ± 0.023	0.620 ± 0.048	0.521 ± 0.024	0.868 ± 0.015
Mental-Bert		0.658 ± 0.031	0.621 ± 0.024	0.701 ± 0.038	0.916 ± 0.007
GPT-5.4		0.658 ± 0.010	0.553 ± 0.027	0.810 ± 0.018	-
Qwen3.5-Plus		0.449 ± 0.010	0.294 ± 0.021	0.943 ± 0.035	-
PAT (BIG5)		**0.665** ± 0.009	0.640 ± 0.026	0.675 ± 0.033	**0.926** ± 0.002
PAT (MBTI)		0.657 ± 0.014	0.631 ± 0.037	0.699 ± 0.030	0.916 ± 0.004
HAN	SWDD	0.824 ± 0.016	0.845 ± 0.005	0.804 ± 0.003	0.966 ± 0.001
DeCapsNet		0.843 ± 0.002	0.878 ± 0.005	0.810 ± 0.005	0.950 ± 0.009
E2LPS		0.851 ± 0.025	0.848 ± 0.015	0.857 ± 0.063	0.978 ± 0.007
Mental-Bert		0.846 ± 0.015	0.844 ± 0.020	0.849 ± 0.026	0.969 ± 0.001
GPT-5.4		0.712 ± 0.011	0.650 ± 0.021	0.788 ± 0.024	-
Qwen3.5-Plus		0.615 ± 0.009	0.493 ± 0.014	0.818 ± 0.028	-
PAT (BIG5)		**0.866** ± 0.013	0.875 ± 0.026	0.859 ± 0.021	**0.980** ± 0.002
PAT (MBTI)		0.852 ± 0.010	0.878 ± 0.018	0.828 ± 0.024	0.974 ± 0.003

Table 1: Main results of experiments in the user-level depression detection.

set is selected for final evaluation. To ensure robustness, each model is trained with five different random seeds. All experiments are conducted on an NVIDIA RTX 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.
- **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. We use **Chinese Mental-BERT** (Zhai et al., 2024) for SWDD.
- **E2LPS** (Wang et al., 2025): This approach achieves end-to-end trainability by using a straight-through technique.
- **LLM**: We perform depression detection via prompting a large language model with the Chain-of-Thought (COT) (Wei et al., 2022) and 2-shot examples from the train set. In practice, we adopt `GPT-5.4` and `Qwen3.5-Plus` for testing.

5 Results

5.1 Main Results

Table 1 presents the comprehensive performance comparison between PAT and various baseline methods on the user-level depression detection task. Experiments demonstrate that PAT (BIG5) achieves state-of-the-art performance across all three benchmarks. Notably, while Mental-BERT, GPT-5.4, and PAT establish substantial leads on the smaller eRisk2017 and eRisk2018 datasets, such superiority narrows or flips on the massive SWDD dataset. This shift suggests that when data is scarce, pre-existing parameterized knowledge plays a guiding role because the data is insufficient for model

optimization. Additionally, since the MBTI setting underperforms the BIG5 setting, we introduce an auxiliary human evaluation on augmented data quality in Appendix A.3.

5.2 Ablation Study

We perform an ablation study for the user-level task under the following configurations: (1) **w/o T**: removing the transfer learning module and employing the original pre-trained model parameters; (2) **without PA**: eliminating data augmentation. (3) **w/o Scorer**: conducting post screening based on the semantic similarity to symptom descriptions instead of our custom scorer. (4) **w/o FGCL**: removing the fine-grained contrastive learning. The results are shown in Table 2. The removal of any single component consistently causes a performance drop, thereby demonstrating the validity and necessity of each individual module in PAT.

Method	eRisk2017 F1	eRisk2018 F1	SWDD F1
PAT	0.640	0.665	0.866
w/o T	0.597	0.593	0.834
w/o Scorer	0.625	0.649	0.812
w/o PA	0.611	0.641	0.843
w/o FGCL	0.631	0.653	0.854

Table 2: The results of the ablation study

5.3 Plug-in Study

The PAT framework can be seamlessly integrated into other baseline methods as a plug-and-play module. Specifically, this involves incorporating our scorer-based filtering during the post-screening stage and applying the transferred parameters from the scorer. We extend PAT to previous approaches; notably, due to their inherent architectural characteristics, Mental-BERT and GPT-5.4 only implement the scorer-based filtering module. As shown in Figure 3, all methods achieve substantial performance gains after incorporating the components of the PAT framework. Remarkably, empowered by our filtering module, GPT-5.4 significantly outperforms all small-model-based competitors on the eRisk2017 and eRisk2018 datasets. This leap strongly validates the precision of our scorer and the efficacy of our filtering strategy.

5.4 Case Study

A detailed user case analysis is provided in Figure 4. The PAT framework not only flags this

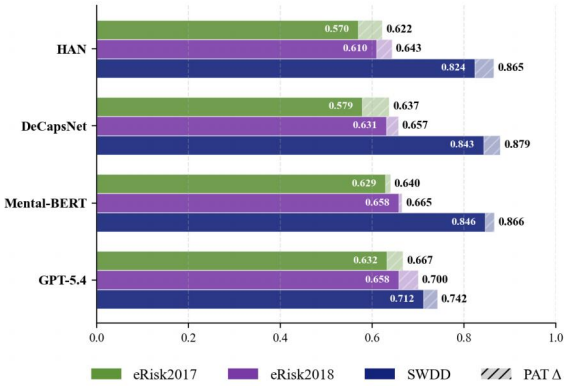


Figure 3: Experimental results (F1 score) of PAT as a plug-in module on baseline methods.

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.6.

Method	Data	Acc	F1
BaseModel	ORI	0.725	0.713
BaseModel	AUG	0.733	0.719
BaseModel + SCL	AUG	0.732	0.717
BaseModel + FGCL	AUG	0.751	0.738

Table 3: Results of post-level depression detection. ORI and AUG denote the original dataset and the augmented dataset, respectively. SCL and FGCL represent Supervised Contrastive Learning and Fine-grained Contrastive Learning.

5.5 Why We Need Fine-grained Contrastive Learning?

During post-level model training, we find that simply adding augmented data yields marginal gains (+0.76% Acc), while further incorporating supervised contrastive learning (SCL) stagnates or even degrades performance. This originates from the overly restrictive assumption in standard SCL, which treats all same-label samples as absolute positive pairs. However, instances sharing the same label are not necessarily identical in the semantic

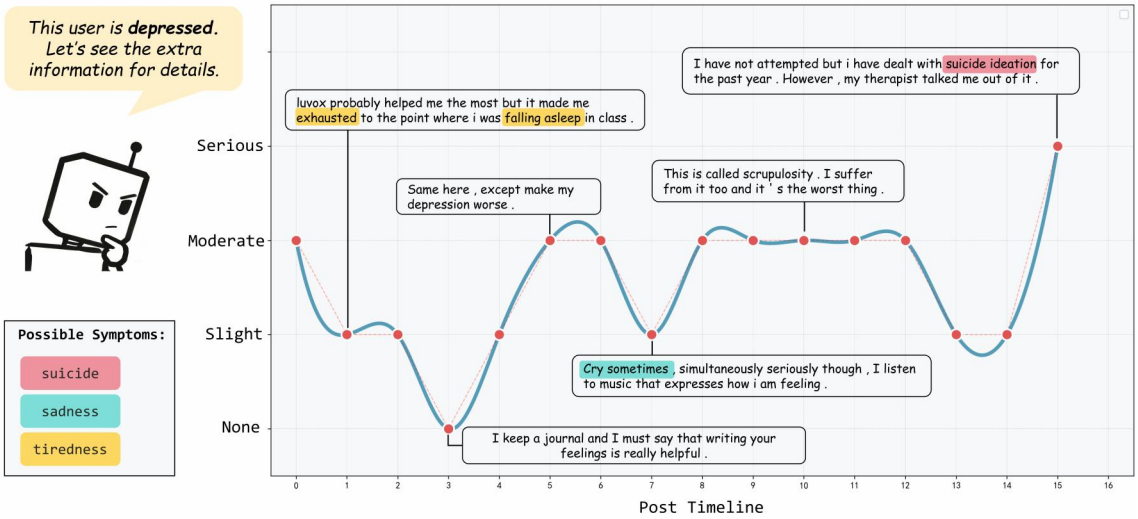
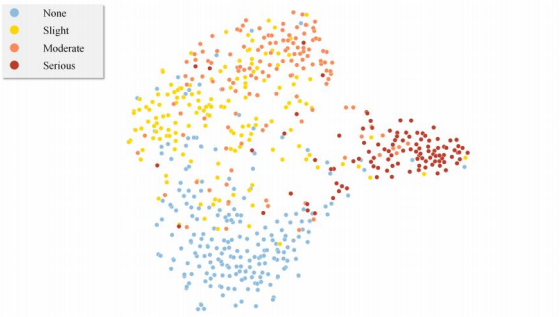


Figure 4: A case study demonstrating the comprehensive interpretability of the PAT framework.

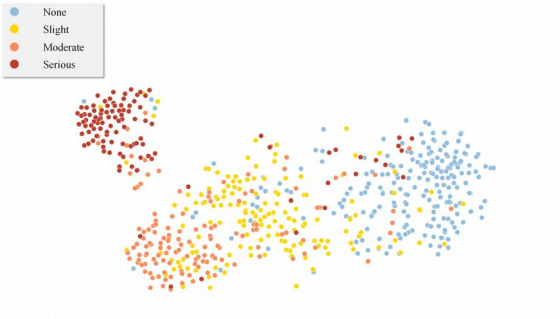
space. An intuitive example is that posts within the same depression severity level can exhibit completely distinct types of symptoms.

Consequently, we propose fine-grained contrastive learning (FGCL). FGCL treats direct augmented counterparts as strong positives like unsupervised contrastive learning, while treating other same-label instances as weaker positives weighted by their relative spatial distances. As shown in Table 3, FGCL triggers a performance increase (+2.62% Acc).

We further visualize the embedding space of contrastive learning on the post-level dataset, as shown in Figure 5. Although supervised contrastive learning can achieve clear separation among the four class clusters, the distances from the "None", "Slight", and "Moderate" to the "Serious" 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.



(a) Trained with Supervised Contrastive Learning



(b) Trained with Fine-Grained Contrastive Learning

Figure 5: Visualization of the embedding space of different contrastive learning methods.

6 Conclusion

In this paper, we present PAT, a personality-driven augmentation and transfer learning framework that effectively addresses existing bottlenecks in user-level depression detection. To alleviate the scarcity of post-level data, we propose personality-driven augmentation and fine-grained contrastive learning techniques specifically tailored for optimizing post-level tasks. This enables the core encoder component to be robustly optimized before being transferred to the user-level model.

Empirical results demonstrate PAT's superior performance and its unique capability to serve as a plug-and-play module for other methods. Moreover, PAT provides comprehensive interpretability by tracing post-level mental-state trajectories and highlighting key symptoms, which greatly enhances the trustworthiness of its predictions and offers actionable insights for clinical 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.

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.

Further Exploration of Fine-Grained Contrastive Learning: While we have demonstrated the effectiveness of fine-grained contrastive learning through empirical results and space visualization, a rigorous mathematical proof is still absent. We plan to explore its theoretical correctness in future work.

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.

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.

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 pre-training 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 used by large language models in the user-level task is shown in Figure 6. The prompt template adopted for the personality-driven augmentation is shown in Figure 7.

Task:
You are a professional mental health analysis assistant. Your task is to analyze a user's social media posts and determine whether they show signs of depression.

You must reason step by step:
1. Symptom Identification: Identify phrases indicating depression symptoms (anhedonia, sleep changes, fatigue, appetite changes, guilt, concentration issues, psychomotor changes, suicidal ideation, etc.)
2. Clinical Mapping: Map symptoms to PHQ-9 / BDI-II clinical domains
3. Severity Assessment: Distinguish clinical expressions from casual language. Clinical expressions reflect sustained patterns; casual language is transient exaggeration
4. Final Judgment: Conclude whether the user is depressed or not

Below are some examples of analyzed users. Use them as reference for your reasoning and output format.

Few-Shot Examples:
[POSITIVE EXAMPLES 1]
[POSITIVE EXAMPLES 2]
[NEGATIVE EXAMPLES 1]
[NEGATIVE EXAMPLES 2]

Now analyze the following user. Output your reasoning and final diagnosis on the last line as either 'depressed' or 'not depressed'.

Input:
[USER POSTS]

Figure 6: The prompt used by large language model methods in the user-level task

A.2 Dataset Statistics

The statistics of the post-level and user-level datasets are shown in Table 4 and Table 7, respectively. Specifically, 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. To adapt the dataset for our regression setting, we quantify the original four-class categorical labels into numerical values $r \in \{0, 1, 2, 3\}$.

	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.

The data augmentation is performed exclusively

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.

BIG5

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.
- Ensure the rewritten sentences are natural, fluent, and conform to everyday language habits.

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 Format:

- Return only the rewritten sentences, no explanations or additional notes. Output for all 10 cases, one per line. For example, the first line should be "High extraversion: [Rewrite for High Extraversion]"

User Input:
[POST TEXT]

MBTI

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 **MBTI** 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.

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.
- Ensure the rewritten sentences are natural, fluent, and conform to everyday language habits.

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 Format:

- Return only the rewritten sentences, no explanations or additional notes. Output for all 8 cases, one per line. For example, the first line should be "Extroverted: [Rewrite for Extroverted]"

User Input:
[POST TEXT]

Figure 7: 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.

on the train set of the post-level dataset. Consequently, the Big Five yields 48,920 augmented posts (10× train set), while the MBTI results in 39,136 posts (8× train set). Concrete examples of these augmented posts are provided in Table 6.

A.3 Assessment of Augmented Data

From the post-level four-class dataset, we selected 25 samples per category, totaling 100 original posts. Consequently, the data augmentation based on the Big Five produced 1,000 augmented sentences (100 posts × 10 traits) per post, and the one based on MBTI generated 800 sentences (100 posts × 8 traits). The evaluation metrics comprise:

- **Semantic Consistency** (SemCons): Whether the augmented sample preserves the depression severity of the original post.
- **Stylistic Consistency** (StyCons): Whether the sample manifests the specified personality traits.

The two dimensions were scored on a 5-point Likert scale, ranging from 1 ("Highly Inconsistent") to 5 ("Highly Consistent"). Three psychology master's students with professional backgrounds participated in this assessment. The final dataset quality scores are summarized in Table 5, where each final score represents the average across the three raters. The results demonstrate the high quality of our data augmentation, with scores for both metrics reaching approximately 4.0 to 5.0. Notably, the Big Five-driven augmentation consistently outperformed the MBTI variant in terms of quality.

Type	SemCons	StyCons
BIG5-0	4.46 ± 0.92	4.86 ± 0.13
BIG5-1	4.32 ± 0.79	4.74 ± 0.32
BIG5-2	4.31 ± 0.91	4.70 ± 0.38
BIG5-3	4.38 ± 1.03	4.68 ± 0.37
BIG5-All	$\mathbf{4.36 \pm 0.91}$	$\mathbf{4.74 \pm 0.31}$
MBTI-0	3.93 ± 1.72	4.47 ± 0.63
MBTI-1	4.34 ± 0.75	4.49 ± 0.52
MBTI-2	4.49 ± 0.47	4.40 ± 0.57
MBTI-3	4.22 ± 0.88	4.27 ± 0.73
MBTI-All	$\mathbf{4.25 \pm 1.00}$	$\mathbf{4.40 \pm 0.63}$

Table 5: Data quality assessment. BIG5-n denotes the augmented sentences derived from original posts labeled with a depression severity of n using the Big Five-driven scheme. BIG5-All represents the complete set of all augmented sentences generated via the Big Five framework

A.4 Symptom Descriptions

Following Wang et al. (2025), we convert the 21 questions from BDI-II into first-person symptom descriptions for post screening. We choose this formulation to homogenize the four distinct severity choices of BDI-II into a single standardized expression, while ensuring that the brief, first-person

From	Trait	Content
Original	*	I'm always the one comforting others, so why is no one comforting me?
BIG5	High Extraversion	I'm always cheering my friends up, so why does no one care about how I feel?
	Low Extraversion	Comforting others is so exhausting; I need to be cared for too.
	High Agreeableness	I am willing to listen to everyone's troubles, but I also hope someone can understand my feelings.
	Low Agreeableness	Why is it always me comforting others? Who cares about my pain?
	High Conscientiousness	I'm used to helping others solve their problems, but I should also learn to seek support for myself.
	Low Conscientiousness	I'll just half-heartedly comfort others; no one cares about me anyway.
	High Neuroticism	I try so hard to comfort everyone, but my own pain feels like a black hole that nobody sees!
	Low Neuroticism	Helping others is a good thing, but occasionally I need to be cared for too.
	High Openness	I am like a lighthouse that is always lit, yet I can never find a harbor of my own.
	Low Openness	I've helped so many people, but when I'm sad, nobody helps me.
MBTI	Extroverted	My friends always come to me to vent, but when I need comfort, nobody is there...
	Introverted	Always listening to other people's troubles, but nobody is willing to listen to my own.
	Sensing	I comforted three friends again today, yet my message list is still empty.
	Intuition	I feel like an emotional trash can—filled with everyone else's pain, but nobody ever comes to empty it.
	Thinking	Analytically speaking, a one-way emotional support relationship is truly exhausting.
	Feeling	Clearly I need to be cared for too, yet I always end up playing the role of the comforter.
	Judging	We agreed to support each other, but in the end, I'm the only one keeping that promise.
	Perceiving	I'm down to chat with anyone who reaches out, but when I need it, I can't find a single soul to talk to.

Table 6: Augmented data examples. Shown here is the English translation of the original Chinese data.

	SWDD			eRisk 2017			eRisk 2018		
	Train	Test	Val	Train	Test	Val	Train	Test	Val
No. of Users (Dep)	2,227	742	742	81	27	27	128	43	43
No. of Users (Con)	11,716	3,905	3,905	451	151	150	896	299	298
No. of Total Users	13,943	4,647	4,647	532	178	177	1,024	342	341

Table 7: Label Distribution and Data Split of the User-Level Datasets.

phrasing closely mirrors the stylistic regularities of social media data. All symptom descriptions are given in Table 8.

A.5 Parameter Sensitivity Analysis of K

We investigate the relationship between the number of retained posts K in the post screening phase and the model performance, with the results illustrated in Figure 8. It can be observed that the model performance generally peaks at $K = 16$, whereas both larger and smaller values of K lead to performance degradation. We attribute this to the fact that an excessive number of posts introduces redundant noise, while a deficient number results in insufficient valid information.

A.6 More Cases

A Case of Healthy User Figure 9 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.

	Symptom	Descriptions
1	Sadness	I feel sad.
2	Pessimism	I am discouraged about my future.
3	Past Failure	I always fail.
4	Loss of Pleasure	I don't get pleasure from things.
5	Guilty Feelings	I feel quite guilty.
6	Punishment Feelings	I expected to be punished.
7	Self-Dislikes	I am disappointed in myself.
8	Self-Criticalness	I always criticize myself for my faults.
9	Suicide	I have thoughts of killing myself.
10	Crying	I always cry.
11	Agitation	I am hard to stay still.
12	Loss of Interest	It's hard to get interested in things.
13	Indecisiveness	I have trouble making decisions.
14	Worthlessness	I feel worthless.
15	Loss of Energy	I don't have energy to do things.
16	Sleep Question	I have changes in my sleeping pattern.
17	Irritability	I am always irritable.
18	Changes in Appetite	I have changes in my appetite.
19	Concentration Difficulty	I feel hard to concentrate on things.
20	Tiredness	I am too tired to do things.
21	Loss of Interest in Sex	I have lost my interest in sex.

Table 8: Symptom descriptions adopted for screening and the symptom-attention module.

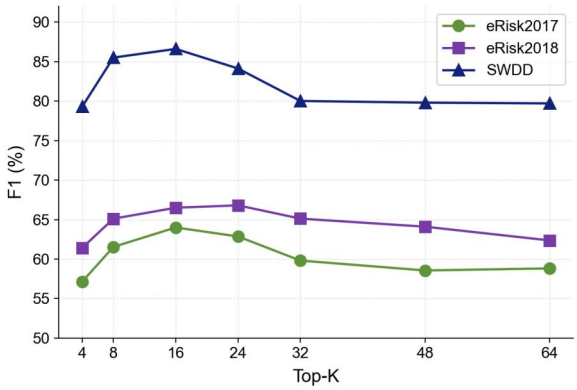


Figure 8: Sensitivity Analysis of Parameter K

Cases of Model Misidentification In this section, we present several error cases for analysis. As shown in Figure 10, 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 misclassified as depressed is presented in Figure 11. 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 should be noted 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 the diagnostic sense, the identification of this case by the model could still be considered meaningful from a risk-aware perspective.

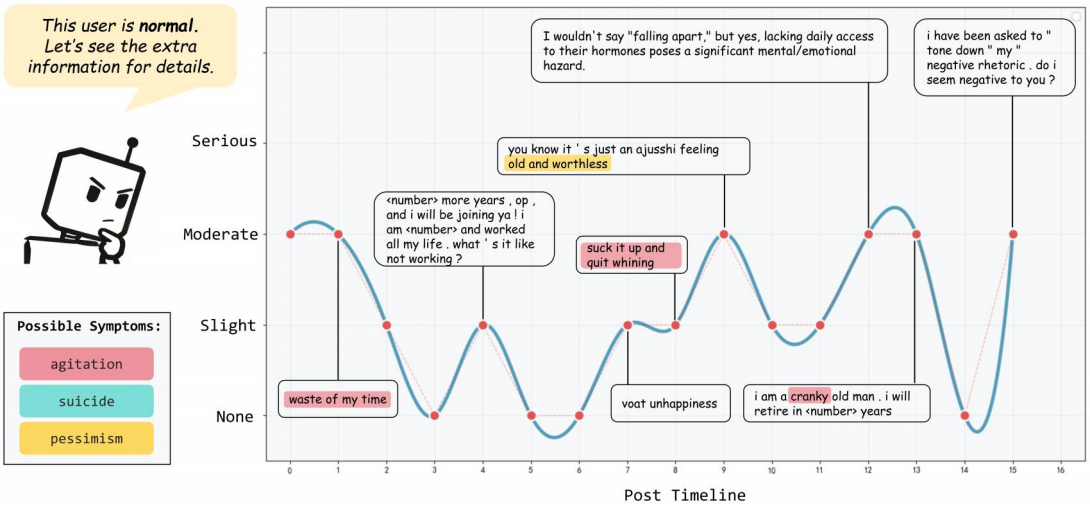


Figure 9: A case of a normal user.

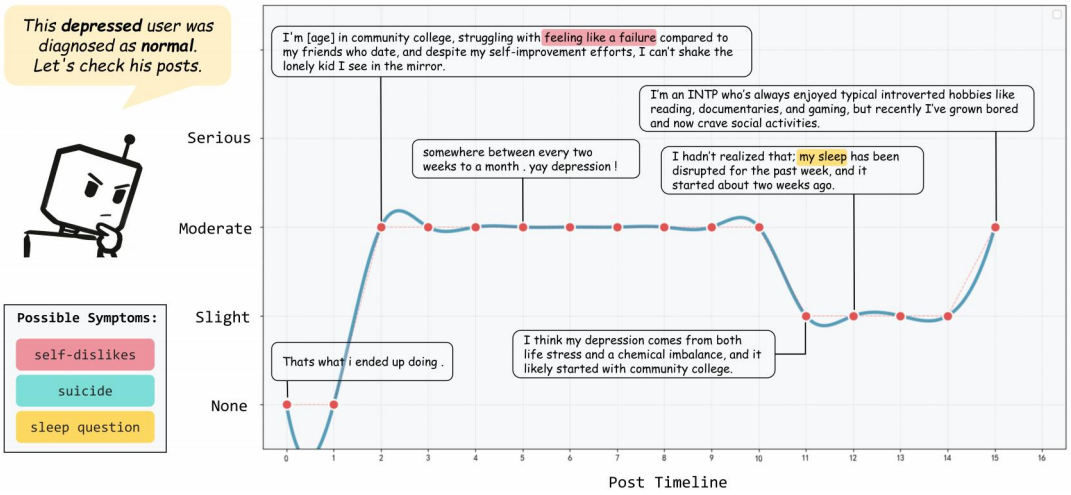


Figure 10: A case of a depressed user that was missed in detection.

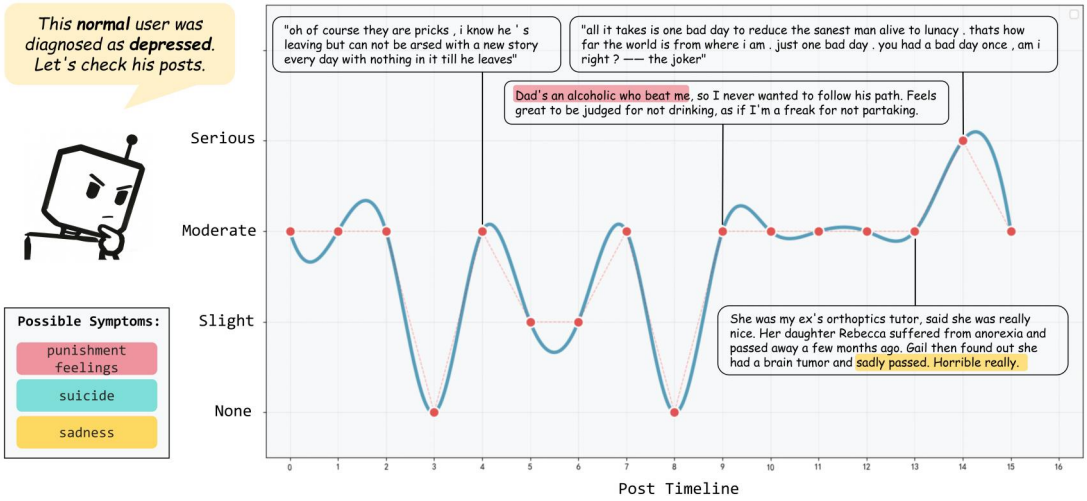


Figure 11: A case of a normal user misdiagnosed with depression.