ALERT: Adversarial Learning Enhanced Stability-aware Routing Transformer for Adaptive Depression Detection

Liangyi Kang[1], **Wei Hua**[1,2], **Yan Yang**[1], **Jie Liu**[*1,2,3], **Dan Ye**[*1, 2, 3]
[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3]Nanjing Institute of Software Technology, Nanjing, China
{kangliangyi15, huawei22, yy, ljie, yedan}@otcaix.iscas.ac.cn

Abstract

Detecting depression through social media is a complex task, as noisy user-generated content creates significant interference between persistent depressive patterns and transient emotional expressions. Two main challenges arise: First, negative mood indicators are not exclusive to depressed individuals, making it difficult to distinguish between pathological symptoms and situational emotional variations. Second, existing static models fail to adapt to diverse user expression styles and effectively filter out confounding noise from posts by non-depressed individuals. This results in conventional approaches either overfitting to superficial emotional cues or overlooking subtle long-term symptom progression. To address these issues, we propose the **A**dversarial **L**earning **E**nhanced Stability-aware **R**outing **T**ransformer for Adaptive Depression Detection(**ALERT**), a novel framework integrating adaptive attention routing and adversarial learning to enhance robustness against confounding mood signals. Specifically, ALERT employs a stability-aware dynamic routing mechanism to annotate user-specific mood valence trends, providing a structured representation of affective progression over time. An adversarial learning module then leverages these mood-based representations to distinguish between expressions indicative of persistent depressive mood and variations in situational mood states, ensuring adaptability to diverse user behaviors. Experimental results on public social media datasets demonstrate that ALERT outperforms state-of-the-art methods in depression detection, effectively reducing false alarm from transient mood states and improving classification accuracy.

Introduction

Depression is a common mental disorder, primarily characterized by persistent depressed mood, loss of interest, and other symptoms lasting for more than two weeks(Association 2013). It affects over 300 million people worldwide, which imposes profound burdens on individuals and healthcare systems. Tragically, clinical diagnosis relies on structured assessments (e.g., PHQ-9) face high costs, stigma, and resource limitations preventing individuals from seeking help. Social media, where users naturally express their psychological state through daily posts, generates vast longitudinal data reflecting mood and behavior. This capacity to capture mood fluctuation and symptom progression aligns closely with core diagnostic criteria in DSM-5 for depression, offering unprecedented opportunities for passive, large-scale mental health screening(Bucur et al. 2023; Wu et al. 2023; Chen et al. 2023). Automatic social media-based depression screening approaches can assist users in recognizing early warning signs and provide healthcare professionals with valuable insights for assessment.

Yet transforming this digital footprint into actionable clinical insight faces a critical roadblock: the pervasive noise masking true depressive signals. This challenge stems from the fundamental difficulty of distinguishing pathological depression patterns - persistent depressed mood aligned with clinical diagnostic criteria - from situational/transient emotional fluctuations that temporarily mimic symptoms. Consider two archetypal cases from our dataset:

- *Case A (Depressed)*: A user diagnosed with depression posts predominantly negative valence expressions (*e.g., "Another sleepless night... why does everything hurt?"*), but occasionally shares positive content (*e.g., "Lovely coffee with Sam."*).
- *Case B (Non-depressed)*: A mentally healthy user going through temporary stress posts negative valence updates (*e.g., "Failed the exam... my world is ending!"*) after finals, but otherwise maintains neutral/positive content.

This creates a *noise-confusion spectrum* where transient expressions obscure persistent symptoms. Note that DSM-5 describes "depressed mood" with examples like feeling sad, empty, or hopeless. Although depression involves complex affective states, analyzing fine-grained emotions from social media data is inherently difficult and may not be directly related to diagnostic criteria. Therefore, we focus on characterizing the overall valence (positive vs. negative polarity) of users' expressed mood in their posts, as this valence-centric approach balances clinical relevance with practical detectability, capturing a core diagnostic dimension while avoiding ambiguous emotion categorization.

Despite the continuous evolution of depression detection methods, existing technologies remain constrained by their inability to robustly navigate the noise confusion inherent in social media data. Multimodal methods(Cheng and Chen 2022; Bucur et al. 2023), while theoretically comprehensive,

*Corresponding authors.

face insurmountable privacy and accessibility barriers by relying on biometric features that exclude text-centric populations and violate evolving regulatory frameworks. Text-based alternatives that span early sequential models(Baytas et al. 2017), transformer variants(Devlin et al. 2019; Bucur et al. 2023), and specialized domain adaptations(Aragón et al. 2023), still operate in fundamentally static architectures. These methods apply uniform processing weights to all user posts and cannot dynamically adjust the focus of attention based on the persistence and expression specificity of symptoms. Even emerging large language models(Xu et al. 2024), despite their representational power, struggle to capture nuanced longitudinal patterns indicative of clinical depression as effectively as tailored approaches. Collectively, these methods rely on relatively static network architectures or uniform processing of user posts. This makes them difficult to dynamically adapt to diverse user-specific mood expression behaviors and robustly distinguish persistent, indicative depressive signals from transient mood fluctuations or noise, leading to false alarms or missed cases.

To bridge this gap, we propose the **A**dversarial **L**earning **E**nhanced Stability-aware **R**outing **T**ransformer Network (**ALERT**), a novel framework designed for reliable early depression detection in noisy social media environments. Departing from conventional static architectures, ALERT first employs a routing transformer with a stability-aware attention mechanism, which automatically weights daily mood expressions based on their temporal consistency - suppressing volatile situational fluctuations while amplifying persistent patterns aligned with clinical diagnostic windows. Building on this temporally refined representation, ALERT then deploys a dual reconstructor adversarial learning module that operates at the semantic level to differentiate pathological expressions from healthy baselines, explicitly identifying depressive camouflage and situational mimicry. This hierarchical design with stability weighting handling temporal noise and adversarial learning resolving semantic ambiguity enables robust adaptation to diverse user expressions.

The main contributions of our paper are summarized as follows.

- We identify the diversity of depressive patterns across users and, for the first time, introduce a stability-aware dynamic routing Transformer design to model user-specific temporal depressive characteristics for adaptive depression detection.
- We propose a novel adversarial learning strategy that utilizes mood trend labeling as a guiding signal. Using dual reconstructors, our model distinguishes depressive tendencies from general mood noise, improving robustness against misleading expressions.
- We conduct experiments on multiple public social media datasets, stating that our approach outperforms existing depression detection methods. Further analysis highlights the model's ability to effectively mitigate mood noise and improve predicting accuracy.

Related Work

Depression Detection Based on Social Media

Early depression detection based on social media data relies on traditional machine learning models with engineered linguistic features(De Choudhury, Counts, and Horvitz 2013; Preotiuc-Pietro et al. 2015; Reece et al. 2016). Deep learning subsequently introduced sequence models such as T-LSTM(Baytas et al. 2017) to capture temporal dependencies. The advent of the Transformer and pre-trained models like BERT significantly advanced contextual understanding. Further research refined these, leading to specialized variants like Time-enriched Transformers incorporating temporal encodings, and domain-adapted models such as EmoBERTa(Kim and Vossen 2021), MentalBERT(Ji et al. 2022), and DisorBERT(Aragón et al. 2023), which leverage mental health-specific corpora or focus on emotional cues to improve relevance and performance within the domain. More recently, diverse architectures like capsule networks in DeCapsNet(Liu et al. 2024), expert-guided systems like PsyEx(Chen et al. 2023), and knowledge distillation in Mood2Content have been explored. The potential of Large Language Models (LLMs), including general ones like GPT-3.5/GPT-4(OpenAI 2023) and specialized versions like Mental-Alpaca and Mental-Flan-T5(Xu et al. 2024), is also under investigation. Despite these advancements, effectively distinguishing persistent depressive mood patterns from transient negative expressions or confounding mood noise across diverse users remains a significant challenge, particularly for models with static architectures or uniform temporal processing.

Adversarial Learning for Robust Text Classification

Adversarial learning has been widely used in natural language processing to improve model robustness, particularly in tasks involving noisy or ambiguous text. Existing adversarial learning approaches in NLP can be broadly categorized into three types. Adversarial domain adaptation(Dai et al. 2020; He, Zhong, and Pan 2022) aim to improve generalization across different datasets or domains by introducing a domain discriminator that forces the model to learn domain-invariant features. Adversarial perturbation-based methods(Choi et al. 2022; Asl et al. 2024) introduce adversarial examples at the input or feature level to improve model robustness against small perturbations. Adversarial learning for feature disentanglement(Miyato, Dai, and Goodfellow 2017) employ adversarial training to separate useful information from noisy or confounding factors in text. Based on adversarial feature disentanglement, our approach leverages mood trend labeling as an adversarial training signal, ensuring that the model remains robust to misleading mood cues in social media data.

Dynamic Routing Network

Dynamic routing is a conditional computation technique that adaptively activates network components based on input, thereby improving computational efficiency. Prior research has integrated it into various architectures. DeCap-

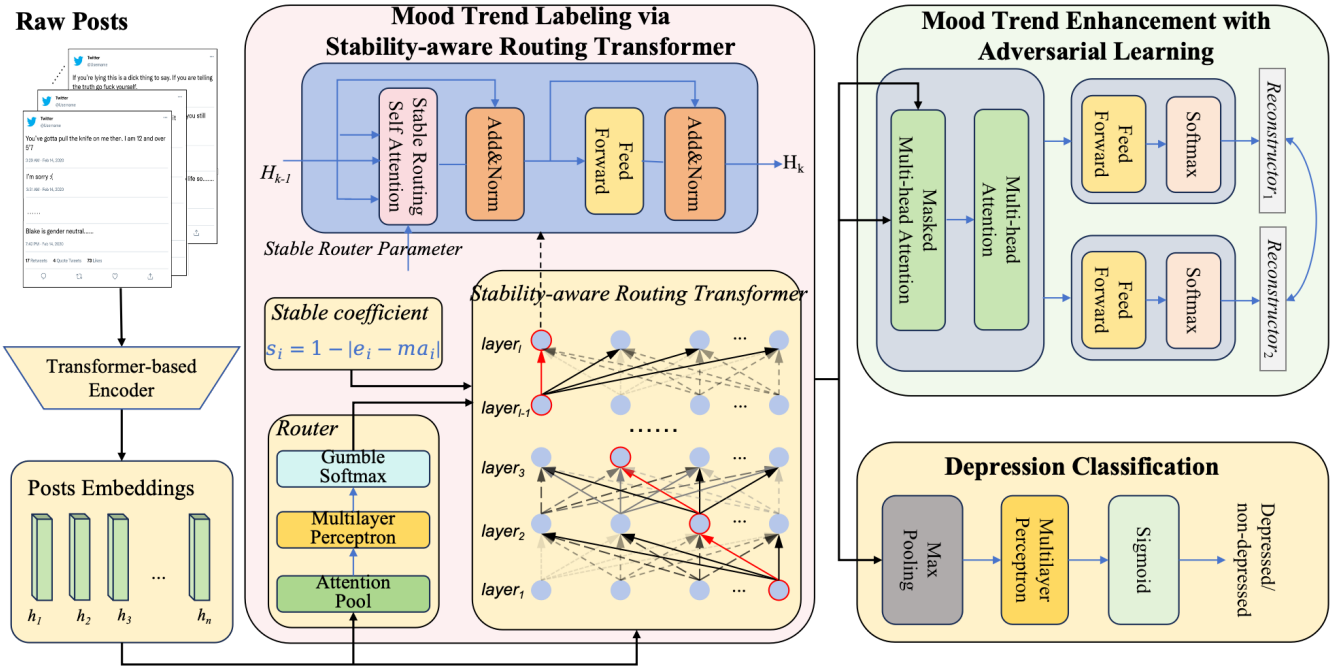

Figure 1: The architecture of Adversarial Learning Enhanced Stability-aware Routing Transformer

sNet(Liu et al. 2024) employs capsules and dynamic routing to enhance depression detection. Switch Transformer(Fedus, Zoph, and Shazeer 2022) incorporates routing in feed-forward layers to reduce computational cost. TRAR(Zhou et al. 2021) integrates it into the self-attention mechanism, strengthening attention capacity without significant overhead. In this work, we extend dynamic routing to user-aware embeddings to handle expressive diversity in social media.

Methods

Problem Formulation

The objective of this study is to detect depression based on the posts on public social media of users. For a given user, the input is defined as $I = \{T_1, T_2, T_3, \ldots, T_n\}$, where n is the total number of days. For posts in i-th day, the data consists of $T_i = \{t_1, t_2, t_3, \ldots, t_q\}$, where q is the number of daily posts. Ultimately, the depression detection task can be formulated as a binary classification problem with a unique output label $y \in \{0, 1\}$, where $y = 1$ indicates depressed and $y = 0$ indicates non-depressed.

Overview

As depicted in Figure 1, ALERT comprises four key components. 1) Posts encoding: embeds raw post sequences via Transformer networks. 2) Mood Trend Labeling via Stability-aware Routing Transformer: dynamically weights posts by mood valence using attention routing, capturing persistent trajectories while filtering transient noise. 3) Mood Trend Enhancement with Adversarial Learning: employs dual reconstructors with mood-guided adversarial learning to distinguish pathological patterns from situational fluctuations in semantic level. This module functions exclusively during the model training stage to improve features robustness. 4) Depression Classification: uses routing-refined features for depression detection.

Posts Encoding

For the posts of i-th day T_i, we concatenate them into a single sequence and separate with [SEP] tokens. The posts of each day T_i are then encoded into a single embedding vector using a pre-trained transformer model as

$$h_i = TF(T_i) \in R^d, \tag{1}$$

where h_i is the daily embedding and d is the hidden size. Then $H = \{h_1, h_2, ..., h_n\}$ inputs following modules.

Mood Trend Labeling via Stability-aware Routing Transformer

Previous approaches often model depressive features in a static manner, limiting their ability to adapt to the diverse expression patterns of different users. To dynamically identify persistent mood trends of users, we employ a Stability-aware Routing Transformer that dynamically selects attention spans for different posts, which focuses on critical and reliable mood cues while filtering transient noise.

Dynamic Attention Routing The Attention Routing is responsible for dynamically assigning mood trend labels to each post. At each layer of routing transformer, the standard attention score is computed as

$$A = \frac{(W_q H) \cdot (W_k H)^T}{\sqrt{d}}, \tag{2}$$

where W_q and W_k are learnable projection matrices. A routing mask AR then selects salient mood expressions, which dynamically adjusts attention scores based on daily mood intensity.

$$AR = \sigma_g(MLP(AttentionPool(H))), \tag{3}$$

where σ_g is Gumbel Softmax with temperature t, MLP is a multilayer perceptron, and $AttentionPool$ aggregates post embeddings.

To enforce temporal locality, we apply a sliding-window adjacency mask M (inspired by (Zhou et al. 2021)). It ensures that posting of each day primarily interacts with its most mood-related neighbors, preventing excessive influence from irrelevant elements. For n days, window w has size $n - 2(w - 1)$, generating binary vectors that restrict attention to local neighbors (e.g., $n = 5$ yields $m_1 = \{1, 1, 1, 0, 0\}$, $m_2 = \{0, 1, 1, 1, 0\}$, etc.). These vectors m collectively form the final adjacency mask by stacking for n times, $M = [m_1, m_2, \ldots, m_n]$.

Stability-aware Modulation To prioritize persistent and reliable mood patterns, we introduce a stability coefficient s_i that quantifies the temporal consistency of daily mood expressions.

$$s_i = 1 - |e_i - ma_i|, \tag{4}$$

where e_i is the day's mood valence from h_i via a linear projection followed by tanh activation, ma_i is its 7-day moving average. This quantifies the deviation from local trends in a half-diagnostic window, aligning with DSM-5's 2-week diagnostic window ($s_i \approx 1$ indicates pathological consistency).

This coefficient modulates the attention weights, amplifying the influence of stable emotional expressions while suppressing volatile fluctuations, ensuring the model focuses on symptom-persistent patterns. The modulated attention becomes:

$$A^{\text{stable}} = (A \cdot M) \odot \mathbf{s}, \tag{5}$$

$$\alpha = \frac{\exp(A^{\text{stable}})}{\sum \exp(A^{\text{stable}})} \cdot AR_j, \tag{6}$$

where $\odot$ denotes element-wise scaling by stability scores.

Mood Trend Representation For each layer l in the Routing Transformer, the mood trend representation is computed as:

$$H_l = \sum_{i=1}^{n} \alpha_i H_{l-1}. \tag{7}$$

The final layer representation is denoted as Z. This representation serves as input for adversarial learning, allowing the model to refine depression detection based on structured mood trends rather than isolated mood fluctuations.

Mood Trend Enhancement with Adversarial Learning

To distinguish persistent depressive patterns from transient mood fluctuations, we introduce a dual-reconstructor adversarial learning module guided by mood trends. This auxiliary task improves robustness against confounding expressions in semantic level through adversarial interaction.

Reconstructor 1: The primary role of Reconstructor 1 is to predict and reconstruct the original content of a user's posts, thereby improving the model's understanding of contextual information and typical user expression patterns under a given mood trend. It operates as:

$$Reconstructor1 = \mathcal{F}(Z, P_M), \tag{8}$$

where $\mathcal{F}$ is a Transformer-based decoder. It takes mood trend features Z and masked daily posts P_M as input. By attempting to fill in the masked content from learned mood trend Z, Reconstructor 1 is forced to learn the general patterns and stylistic elements present in the user's typical content flow when exhibiting that particular mood trend. We mask the three days with highest routing weights and include two randomly selected unmasked days to focus on key mood expressions while promoting generalization. The loss for Reconstructor 1, L_{R1}, is thus a weighted combination of losses from masked and unmasked predictions:

$$L_{R1} = \lambda L_{\text{mask1}} + (1 - \lambda) L_{\text{unmask1}}, \tag{9}$$

where λ is a hyperparameter balancing the two types of predictions.

Reconstructor 2: Operating antagonistically, Reconstructor 2 strives to distinguish Reconstructor 1's predictions from actual user content. This adversarial task enables the overall model to identify subtle deviations in a user's mood expression patterns that might signify genuine depressive tendencies, rather than just common fluctuations. If Reconstructor 1, guided by a "non-depressed" mood trend Z, generates content that significantly differs from a user's actual highly negative post, Reconstructor 2 learns to flag this discrepancy, thereby helping to differentiate true depressive signals from what might be considered "typical" (non-depressive) mood expressions. The architecture of Reconstructor 2 is identical to Reconstructor 1, but its loss function is designed to incentivize this discriminative capability:

$$\begin{aligned} L_{R2} = & \frac{1}{n_e} \left(\lambda L_{\text{mask2}} + (1 - \lambda) L_{\text{unmask2}}\right) \\ & - (1 - \frac{1}{n_e}) L_{R1}. \end{aligned} \tag{10}$$

Here, n_e is the training epoch. The term $\frac{1}{n_e}$ is used because, in the early stages of training (small n_e), we want both reconstructors to focus on learning the fundamental patterns of mood fluctuations in user content without excessive adversarial pressure. As training progresses, the adversarial component becomes more dominant. Once Reconstructor 2 is adequately trained, it can effectively discern whether the content predicted by Reconstructor 1 genuinely aligns with established, learned mood patterns or if it deviates in a way that might indicate a depressive state.

Through this interplay, the two reconstructors enable the model to identify typical mood fluctuations and, more importantly, to enhance its sensitivity to subtle yet persistent depressive signals often obscured within diverse social media contexts. This entire adversarial learning module functions exclusively during the model training phase to refine and robustify the learned text features, and is disabled during inference.

	STMHD		MultiRedditDep		eRisk2018	
	Dep.	Non-Dep.	Dep.	Non-Dep.	Dep.	Non-Dep.
users	6803	5000	1419	2344	214	1493
posts/user	7767	9845	4752	2547	422	661

Table 1: Datasets statistics.

Depression Classification

For the final classification step, the feature representations Z obtained from the Routing Transformer undergo max-pooling. These pooled features are then processed through a multi-layer perceptron (MLP) followed by a sigmoid activation layer, which classifies the user as either depressed or non-depressed. The classification can be formulated as

$$p = \sigma\left(W \cdot \text{max-pooling}(Z) + b\right), \tag{11}$$

where W is a trainable parameter and b is the bias.

Loss of the classification module is binary cross entropy:

$$L_C = \sum y_i log(p_i), \tag{12}$$

where y_i is the real label and p_i is the probability output of the model.

Overall Loss Function

By combining Eqs.(9, 10, 12), we have the overall loss function of the proposed method.

$$L = \delta L_C + (1-\delta)(L_{R1} + L_{R2}), \tag{13}$$

where δ is a hyper parameter. By minimizing the loss L with the gradient descent method, all trainable parameters can be learned.

Experiments

Datasets

To evaluate our depression detection model, we use three widely available benchmark datasets, each containing users with posts spanning over 14 days. Twitter **STMHD**(Suhavi et al. 2022) is a large-scale user-level dataset focusing on mental health disorders, including depression. **MultiRedditDep**(Uban, Chulvi, and Rosso 2022) is a multimodal dataset designed derived from Reddit. We focus on its text modality, which provides longer and more expressive content. **eRisk2018**(Losada and Crestani 2016) is the dataset comprising user-generated posts and comments from Reddit platforms, structured to distinguish between users diagnosed with depression and those not (control group). For all datasets, we follow the standard data split ratio of 7:2:1 for training, validation, and testing, consistent with prior studies. Table 1 summarizes the detailed statistics of datasets.

Baselines

For a comprehensive evaluation, we compare our proposed method (ALERT) against a diverse set of baseline models. These baselines are grouped into four distinct categories:

- **Sequence Models:** This group includes fundamental deep learning models, **T-LSTM**(Baytas et al. 2017) and standard **Transformer**(Vaswani et al. 2017), that process text sequentially to capture context, without necessarily specializing in the nuances of mental health.
- **BERT and Fine-tuned Variants:** This category comprises models based on the BERT architecture fine-tuned for tasks specifically adapted for emotion or mental health domains, including **BERT**(Devlin et al. 2019), **EmoBERTa**(Kim and Vossen 2021), **MentalBERT**(Ji et al. 2022) and **DisorBERT**(Aragón et al. 2023).
- **Task-Specific Advanced Models for Depression Detection:** This category includes models employing advanced architectures or explicitly integrating specific features relevant to depression detection (symptoms, emotions, time). **DeCapsNet**(Liu et al. 2024) is a capsule network using contrastive learning and dynamic routing to extract depression-related symptom features. **PsyEx**(Chen et al. 2023) designs two-stream model leveraging psychiatric domain knowledge and symptom-based screening mechanisms. **Mood2Content**(Wu et al. 2023) integrates textual content with explicitly extracted mood-related features using knowledge distillation. **Time-enriched Transformers**(Bucur et al. 2023) is a Transformer architecture to model temporal dynamics.
- **Large Language Models (LLMs):** This category encompasses recent large-scale generative and instruction-tuned language models, including some adapted for the mental health domain. **GPT-3.5**(OpenAI 2022), a large language model from OpenAI that is applied via zero-shot classification. **GPT-4**(OpenAI 2023) is an improved version. **Mental-Alpaca**(Xu et al. 2024), an instruction-tuned LLM adapted based on Alpaca for mental health tasks. **Mental-Flan-T5**(Xu et al. 2024), an instruction-tuned model based on Flan-T5, adapted for the mental health domain.

Training setup

According to the depression definition DSM-5, the diagnostic standard for depression requires symptoms such as persistently depressed mood or marked decrease in interest or pleasure for most of the time over at least two consecutive weeks. To align with this criteria, we select 14 consecutive days of user posts as input. To accommodate irregular posting, we extend the window to 21 days requiring 14 active days, with 10 posts/day for sufficient context. The post encoder uses identical BERT-base encoders compared to existing depression detection methods, and trains together with the model. The routing Transformer uses 6 layers and 8 heads, trained for 30 epochs (batch size 16) with Adam optimizer (lr 10^{-7}), dropout 0.5, Gumbel Softmax temperature $t = 10$, and hyperparameters $\lambda = 0.8$, $\delta = 0.7$.

Results

In terms of the evaluation metrics, we follow (Liu et al. 2024) to use F1-score and Accuracy. The results are presented in Table 2 , and we obtain following observations based on model categories:

	STMHD		MultiRedditDep		eRisk2018	
	F1	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
Transformer	69.87	70.49	83.78	87.66	62.48	61.97
T-LSTM	67.41	66.77	83.11	87.27	61.39	58.94
BERT	68.21	69.89	83.01	86.93	64.47	64.18
EmoBERTa	70.12	71.08	84.36	87.94	67.37	66.28
Disorbert	80.64	80.52	87.05	87.26	81.27	82.67
Mentalbert	70.68	71.79	86.92	87.33	79.35	82.54
DeCapsNet	69.79	70.23	85.36	84.77	79.00	73.24
PsyEx	72.05	76.51	86.12	87.89	82.12	82.31
Mood2Content	71.12	71.79	80.21	83.17	80.22	82.93
Time-enriched Transformers	73.58	71.93	83.54	83.21	77.61	73.89
GPT3.5	63.29	65.28	76.33	77.96	70.39	71.28
GPT4	68.73	69.09	80.29	81.37	77.98	76.25
Mental-Alpaca	70.74	72.84	84.36	84.83	81.38	82.95
Mental-Flan-T5	71.96	72.38	83.82	84.15	80.19	82.47
ALERT (Ours)	**84.86**	**86.53**	**88.26**	**89.38**	**84.17**	**85.38**

Table 2: The experimental results [%] comparing baseline methods and our method. Models are grouped by type, following specific formatting requests. The best score in each column is highlighted in bold font, and the second-best score is underlined. Results are averaged over 5 runs.

- **Comparison with Sequence Models:** Basic sequence models (Transformer and T-LSTM) generally underperform compared to more specialized architectures. While they capture sequential dependencies, they lack mechanisms to specifically address the nuances of temporal mood shifts or domain-specific language often found in depression-related text, limiting their effectiveness.
- **Comparison with BERT and Fine-tuned Variants:** Fine-tuning BERT improves upon basic sequence models. Further specialization, such as incorporating emotion awareness (EmoBERTa) or domain-specific pretraining (MentalBERT, DisorBERT), yields significant gains. DisorBERT and EmoBERTa achieve second-best on STMHD and MultiRedditDep dataset. This highlights the advantage of leveraging pre-trained models and adapting them to the specific domain or task facets like emotion. However, even these variants might not fully capture the dynamic interplay of features over time that our model addresses.
- **Comparison with Task-Specific Advanced Models:** Models explicitly designed for depression by integrating symptoms (DeCapsNet, PsyEx), mood (Mood2Content), or time (Time-enriched Transformers) show strong performance, often outperforming generic or BERT-based approaches on specific metrics. PsyEx, leveraging symptom knowledge, achieves the second-best F1 on eRisk2018.These results underscore the importance of incorporating depression-relevant features. However, integrating these features within static architectures might limit adaptability compared to our dynamic approach. DeCapsNet and Mood2Content show reasonable but not top-tier performance across the board.
- **Comparison with Large Language Models (LLMs):** LLMs present a mixed picture. While models instruction-tuned for the mental health domain achieve respectable results, with Mental-Alpaca obtaining the second-best accuracy on eRisk2018, they generally do not surpass the best-performing specialized models. General-purpose LLMs like GPT-3.5 and GPT-4 lag significantly, suggesting that scale alone is insufficient without task-specific adaptation or fine-tuning for this nuanced classification problem.
- **Performance of Proposed Method (ALERT):** Our proposed method consistently achieves state-of-the-art results, securing the highest F1-score and Accuracy on all three datasets. ALERT surpasses the second-best F1 scores by approximately 4.22% on STMHD (vs. Disorbert), 1.21% on MultiRedditDep (vs. DisorBERT), and 2.05% on eRisk2018 (vs. PsyEx). We attribute ALERT's strong performance to its ability to dynamically adapt to user-specific mood trends via the routing transformer and effectively distinguish genuine depressive patterns from noise using the guided adversarial learning.

Model	F1	Acc.	FP	FN
Ours	84.86	86.53	14.72	9.04
- w/o Reconstructor2	82.27	82.35	15.78	13.13
- w/o Reconstructor1+2	80.48	81.29	18.12	15.02
- w/o Stability coefficient	84.29	85.54	15.02	9.86
- w/o Routing+Reconstructor	79.23	79.48	20.03	17.89

Table 3: The experimental results [%] of ablation test on STMHD. FP and FN are false positive and negative scores.

Ablation Study

To validate the effectiveness of various modules in our model, we conduct an ablation study by systematically removing different components. Since reconstructor2 depends on reconstructor1, and the entire adversarial reconstructor

Number of Days	F1	Accuracy
7	78.33	77.29
11	82.07	83.24
14	84.86	86.53
17	85.93	87.04
21	86.19	87.28

Table 4: The experimental results [%] of posts with different total days on STMHD dataset.

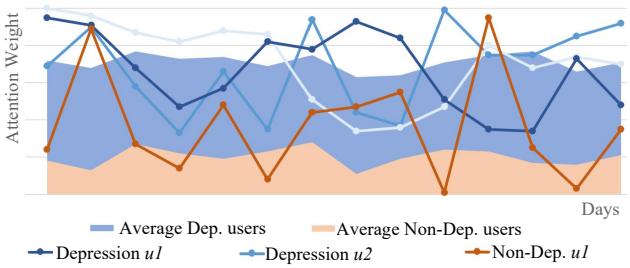

Figure 2: Mood trajectories of average and example users.

module relies on the mask matrix provided by the Routing Transformer module, we performe four ablation experiments: removing reconstructor1, removing both reconstructor1 and reconstructor2, removing stability coefficient and both the routing and reconstructor modules. The results in Table 3 show that each removal leads to a decrease in both F1-score and accuracy, indicating that the two-phase reconstruction mechanism effectively prevents the model from overfitting to superficial mood cues.

Results indicate that non-depressed users frequently express transient negativity during stressful events (primary source of False Positives), while depressed users predominantly maintain negative expressions (contributing to low False Negatives). ALERT significantly mitigates both error types, which reduces FP from situational distress and FN from symptom-masking behaviors. Ablation reveals dual-Reconstructor's critical role in false alarm suppression, as it can model the contextualize intense emotional language semantically.

The Effect of Posts Number

We further investigate the impact of varying the number of days that user posts on the social media. The results on STMHD dataset are shown in Table 4. It indicates that increasing the number of total days for posts leads to improvements in both F1-score and accuracy. Expanding the total number of days from 7 to 14 yields noticeable gains. While performance continues to improve from 14 to 21 days, the rate of improvement slightly decelerates. This confirms ALERT's ability for early screening by detecting patterns that align with DSM-5 diagnostic threshold (two weeks), providing a critical foundation for long-term monitoring.

Case study

Figure 2 visualizes mood trend trajectories via routing attention weights, contrasting average group trends (shaded areas) with specific daily patterns for three depressed and one non-depressed user (lines). These weights combine attention scores and dynamic routing coefficients. Although the average (shaded) trajectory for depressed users is consistently higher, reflecting clinical expectations, the individual (line) trajectories reveal significant variability and mood noise. For instance, depressed users show occasional dips, and the non-depressed user exhibits transient spikes in depressed mood expression, deviating from their respective group averages. This underscores the challenge of confounding signals and validates the need for ALERT's Adversarial Learning to discern persistent depressive patterns from such noise.

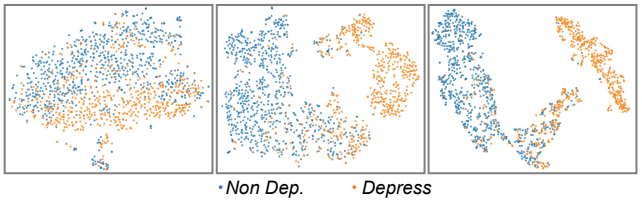

Figure 3: t-SNE visualization of Feature spaces on STMHD dataset, from left to right are based on raw embedding, stability-aware routing embedding and adversarial learning enhanced embedding.

Figure 3 presents t-SNE visualizations of the feature space evolution during processing stages. Raw embeddings (left) show significant overlap. After stability-aware routing (center), distinct peninsulas begin emerging—depressed samples coalesce into denser formations while maintaining filamentous connections to non-depressed clusters. Finally, the adversarial-enhanced embedding (right) achieves clearest class separation and depressed features organize into compact masses, which confirms ALERT's robustness against common confounding factors. However, there are still residual mixing cases, and it occurs primarily where users employ ambiguous language (e.g., sarcasm) or symptom-masking expressions.

Conclusion

This paper proposes ALERT (Adversarial Learning Enhanced Stability-aware Routing Transformer), a novel framework for adaptive depression early screening based on social media. ALERT integrates dynamic mood trend routing with an adversarial learning mechanism using dual reconstructors. The routing transformer adaptively weights posts by mood persistence across a user's posting history to filter transient noise of different user expressions. The adversarial module leverages these refined representations to distinguish pathological depressive patterns from situational mood fluctuations in semantic level. Extensive validation demonstrates state-of-the-art performance across datasets, with component synergy analysis confirming the framework's robustness against social media noise and effectiveness of reducing false alarm.

Ethical Statement

Our framework is designed to balance practical impact with privacy considerations. It utilizes only publicly accessible, anonymized text data at training, avoiding invasive biometric modalities, which inherently minimizes privacy risk. In practice, ALERT could be deployed as a consent-based plugin on social platforms, performing local early screening to encourage help-seeking without exposing raw personal data. Technically, ALERT directly addresses the key barrier to real-world deployment, false alarms from transient moods, through its stability-aware routing and adversarial modules. Its clinical alignment (DSM-5) and adaptability to diverse user-specific expression styles ensure robustness and reduce algorithmic bias. This provides a scalable and privacy-conscious path toward positive social impact.

Note that, we emphasize the role of ALERT as a responsible screening aid that supports and not replaces clinical judgment, thus mitigating potential harm by ensuring human oversight.

Acknowledgments

This work is supported by The National Science and Technology Innovation 2030 "Neuroscience and Brain-like Research" Major Project, Prospective Clinical Cohort Study on Depression (Grant No.2021ZD0200603), and Basic Research Project of Institute of Software, Chinese Academy of Sciences (Grant No.ISCAS-JCMS-202405).

References

Aragón, M. E.; López-Monroy, A. P.; Gonzalez, L.; Losada, D. E.; and Montes, M. 2023. DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 15305–15318. Association for Computational Linguistics.

Asl, J. R.; Panzade, P.; Blanco, E.; Takabi, D.; and Cai, Z. 2024. RobustSentEmbed: Robust Sentence Embeddings Using Adversarial Self-Supervised Contrastive Learning. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 3795–3809. Association for Computational Linguistics.

Association, A. P. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. American Psychiatric Association.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, 65–74. ACM.

Bucur, A.; Cosma, A.; Rosso, P.; and Dinu, L. P. 2023. It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, volume 13980 of *Lecture Notes in Computer Science*, 200–215. Springer.

Chen, S.; Zhang, Z.; Wu, M.; and Zhu, K. Q. 2023. Detection of Multiple Mental Disorders from Social Media with Two-Stream Psychiatric Experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 9071–9084. Association for Computational Linguistics.

Cheng, J.; and Chen, A. L. P. 2022. Multimodal time-aware attention networks for depression detection. *J. Intell. Inf. Syst.*, 59(2): 319–339.

Choi, S.; Jeong, M.; Han, H.; and Hwang, S. 2022. C2L: Causally Contrastive Learning for Robust Text Classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 10526–10534. AAAI Press.

Dai, Y.; Liu, J.; Ren, X.; and Xu, Z. 2020. Adversarial Training Based Multi-Source Unsupervised Domain Adaptation for Sentiment Analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7618–7625. AAAI Press.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 3267–3276. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318990.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.*, 23: 120:1–120:39.

He, Z.; Zhong, Y.; and Pan, J. 2022. An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition. *Computers in biology and medicine*, 141: 105048.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, 7184–7190. European Language Resources Association.

Kim, T.; and Vossen, P. 2021. EmoBERTa: Speaker-Aware

Emotion Recognition in Conversation with RoBERTa. *CoRR*, abs/2108.12009.

Liu, H.; Li, C.; Zhang, X.; Zhang, F.; Wang, W.; Ma, F.; Chen, H.; Yu, H.; and Zhang, X. 2024. Depression Detection via Capsule Networks with Contrastive Learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, 22231–22239. AAAI Press.

Losada, D. E.; and Crestani, F. 2016. A Test Collection for Research on Depression and Language Use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, 28–39. Springer.

Miyato, T.; Dai, A. M.; and Goodfellow, I. J. 2017. Adversarial Training Methods for Semi-Supervised Text Classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

OpenAI. 2022. OpenAI API - Models (GPT-3.5). https://platform.openai.com/docs/models/gpt-3.5-turbo.

OpenAI. 2023. GPT-4 Technical Report. Technical report, OpenAI.

Preotiuc-Pietro, D.; Eichstaedt, J. C.; Park, G. J.; Sap, M.; Smith, L.; Tobolsky, V.; Schwartz, H. A.; and Ungar, L. H. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych@NAACL-HLT 2015, June 5, 2015, Denver, Colorado, USA*, 21–30. The Association for Computational Linguistics.

Reece, A. G.; Reagan, A. J.; Lix, K. L. M.; Dodds, P. S.; Danforth, C. M.; and Langer, E. J. 2016. Forecasting the onset and course of mental illness with Twitter data. *CoRR*, abs/1608.07740.

Suhavi; Singh, A. K.; Arora, U.; Shrivastava, S.; Singh, A.; Shah, R. R.; and Kumaraguru, P. 2022. Twitter-STMHD: An Extensive User-Level Database of Multiple Mental Health Disorders. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, 1182–1191. AAAI Press.

Uban, A.-S.; Chulvi, B.; and Rosso, P. 2022. *Explainability of Depression Detection on Social Media: From Deep Learning Models to Psychological Interpretations and Multimodality*, 289–320. Cham: Springer International Publishing.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wu, J.; Wu, X.; Hua, Y.; Lin, S.; Zheng, Y.; and Yang, J. 2023. Exploring Social Media for Early Detection of Depression in COVID-19 Patients. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 3968–3977. ACM.

Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J. A.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1): 31:1–31:32.

Zhou, Y.; Ren, T.; Zhu, C.; Sun, X.; Liu, J.; Ding, X.; Xu, M.; and Ji, R. 2021. TRAR: Routing the Attention Spans in Transformer for Visual Question Answering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2054–2064. IEEE.