# End-to-End Learnable Psychiatric Scale Guided Risky Post Screening for Depression Detection on Social Media

**Bichen Wang*** **Yuzhe Zi*** **Yixin Sun** **Yanyan Zhao†** **Bing Qin**
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, Heilongjiang, China
{bichenwang,yuzhezi,yxsun,yyzhao,qinb}@ir.hit.edu.cn

## Abstract

Detecting depression through users' social media posting history is crucial for enabling timely intervention; however, irrelevant content within these posts negatively impacts detection performance. Thus, it is crucial to extract pertinent content from users' complex posting history. Current methods utilize frozen and static screening models, which can miss critical information and limit overall performance due to isolated screening and detection processes. To address these limitations, we propose **E2-LPS** (**E**nd-to-**E**nd **L**earnable **P**sychiatric Scale Guided Risky Post **S**creening Model) for jointly training our screening model, guided by psychiatric scales, alongside the detection model. We employ a straight-through estimator to enable a learnable end-to-end screening process and avoid the non-differentiability of the screening process. Experimental results show that E2-LPS outperforms several strong baseline methods, and qualitative analysis confirms that it better captures users' mental states than others.

## 1 Introduction

According to data from the WHO, depression affects approximately 3.8% of the world's population[1]. In the U.S., nearly 15% of adults experience a major depressive episode in their lifetime (Kessler et al., 2005). In response, global efforts have focused on utilizing social media for detection to mitigate the severe consequences of depression (Zhou et al., 2018; Malhotra and Jindal, 2022). [2]

User-level depression detection aims to detect depression from social media posting history. However, as illustrated in Figure 1, because depression has an episodic nature with symptom-free periods (Ma, 2021), much non-depressive content is generated, making the tracking of every post often unnecessary. The vast volume of user posts further complicates this, necessitating some methods to screen critical content. Current approaches enhance accuracy by capturing relevant mental state information and filtering irrelevant content before detection. For example, Zogan et al. (2021) uses hybrid summarization to filter relevant tweets for detailed information, while Zhang et al. (2022); Liu et al. (2024) employ psychiatric scales to statically screen risky posts indicative of mental states. These methods have achieved some success.

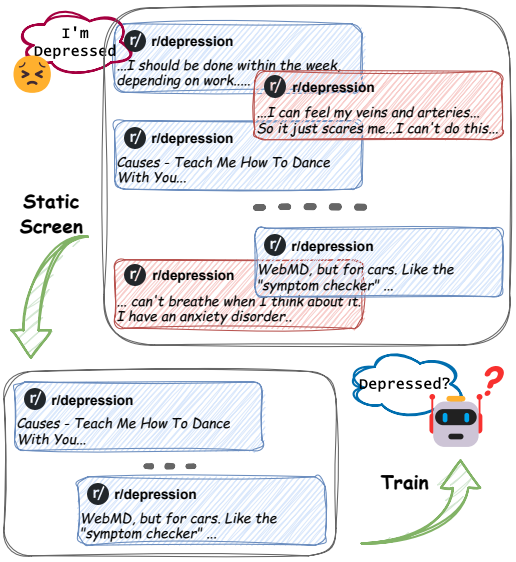


Figure 1: Limitations of static unsupervised screening. User histories contain much irrelevant content, which static screening often includes. Because it lacks feedback from the detection model, this irrelevant content directly trains the model. Training on such noisy data confuses the model and reduces its performance.

Existing methods typically separate screening and detection, relying on the results of static, prior screening processes. While these unsupervised screening approaches are somewhat effective, they

*These authors contributed equally to this work.
†Corresponding author.
[1] https://www.who.int/news-room/fact-sheets/detail/depression
[2] This paper contains content on depression and mental health, which may be sensitive or distressing for some readers.



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4055–4067
November 4-9, 2025 

often lack strong feedback because they don't learn from detection performance. As shown in Figure 1, inaccurate screening results can introduce detection errors, leading to models being trained on inappropriate data and degrading performance. To overcome these limitations, we propose a more flexible, learnable screening process that is trained end-to-end with other components. This approach effectively filters out irrelevant content and captures risky posts most relevant to users' mental health, thereby improving detection performance.

Overall, to address this limitation, we propose E2-LPS (**E**nd-to-**E**nd **L**earnable **P**sychiatric Scale Guided Risky Post **S**creening for Depression Detection on Social Media). E2-LPS is an end-to-end training method that integrates the psychiatric scale-guided screening process with the subsequent detection process. This method enables the screening process to receive direct feedback from overall detection performance. Specifically, E2-LPS enhances depression detection by directing the detection model to focus on risky posts identified as highly relevant by the screening model. Additionally, it produces improved screening results, which can serve as a reference to assist experts in implementing more effective subsequent interventions. Since the direct screening process may not be differentiable, we employ straight-through estimators to parameterize the screening decision, transforming the screening process from a passive, isolated task into an active, supervised learning component.

Extensive experiments, both within- and cross-dataset, demonstrate E2-LPS's superior performance against strong existing baselines in social media depression detection. Furthermore, expert comparisons confirm our screening results are more clinically meaningful and better at identifying daily risky posts of depression. Our contributions are summarized as follows:

- We integrate the screening process and detection process into a joint training framework to solve the previous problem.
- We propose our E2-LPS model to extract information most relevant to users' mental health based on psychiatric scales.
- We conduct experiments across multiple datasets. Comprehensive within-dataset and cross-dataset evaluations demonstrate that our E2-LPS significantly outperforms many strong existing baselines.

## 2 Related Work

### 2.1 User-Level Psychiatric Problem Detection

Modern individuals share their lives on social media, revealing personal activities and characteristics that offer valuable insights for modeling psychiatric states. Computational techniques are increasingly used for analyzing users' mental health states from this data, which studies show can detect disorders like depression, PTSD, and anxiety (Choudhury et al., 2013; Coppersmith et al., 2015). Identifying depression symptoms is a key research focus. CNNs and RNNs have significantly improved depression detection accuracy over traditional methods (Husseini Orabi et al., 2018), as seen in challenges like CLPsych-2015 (Coppersmith et al., 2015). Due to lengthy, often irrelevant post histories, summarization and scale-based screening are employed to extract depression-relevant information (Zogan et al., 2021; Zhang et al., 2022; Liu et al., 2024). While Large Language Models (LLMs) incorporating scales and Chain-of-Thought (COT) reasoning are being explored for mental health detection (Wang et al., 2024a), efficiency constraints still demand simple screening methods for collecting representative posts, especially for resource-limited real-time applications.

### 2.2 Depression Detection

Depression detection often uses data from clinical interviews or social media (Gratch et al., 2014; Salas-Zárate et al., 2022). Diagnostic cues come from self-reports or involvement in depression communities (Ernala et al., 2019). Some studies use visual cues and engagement metrics via VGGNet (Simonyan and Zisserman, 2015; Zogan et al., 2021), while others combine psychiatric features with neural networks, integrating topic features with attention-enhanced pre-trained models (Song et al., 2018; An et al., 2020). Other research uses standardized scales, focusing on metaphor and moral language in diagnosis (Coll-Florit et al., 2021; Han et al., 2022). Recent studies emphasize early detection (Sadeque et al., 2018), proposing psychiatric scale-based risk screening (Zhang et al., 2022) and early detection loss functions (Wang et al., 2024b).

However, these methods usually depend on complete user history or basic static screenings, separating screening from detection. Our approach try to create a learnable dynamic screening mechanism, providing an end-to-end solution, which improves model by refining the information for detection.

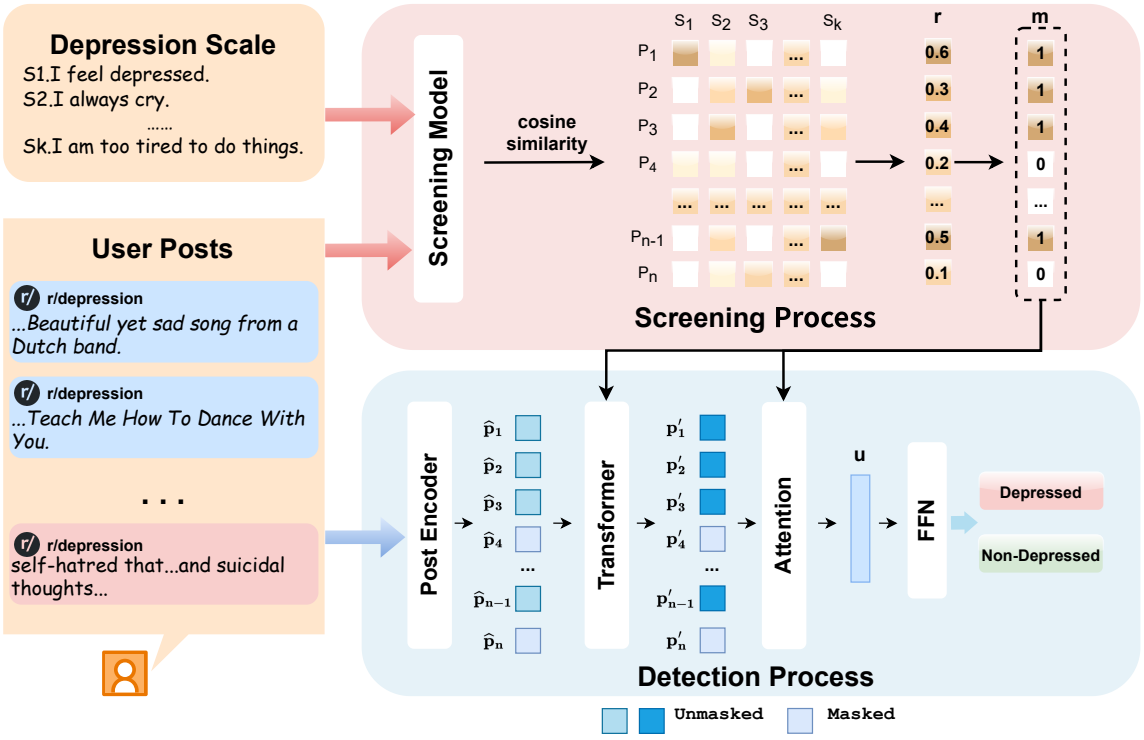


Figure 2: An overview of the E2-LPS method: Unlike traditional depression detection methods, we employ a psychiatric scale-based screening model to identify significant risky posts and utilize an end-to-end framework to optimize the E2-LPS model. During training, E2-LPS allows the screening process to adaptively focus on the most relevant content by receiving feedback signals from the detection process.

## 3 E2-LPS: End-to-End Learnable Psychiatric Scale Guided Risky Post Screening Model

**Problem Formulation:** Given a user $U$ with posts $P = \{p_1, \ldots, p_n\}$, where $p_i$ is the $i$-th post, we aim to detect if the user is depressed based on $P$. This involves predicting a label $y \in \{0, 1\}$, indicating the user's mental health: $y = 0$ (not depressed), or $y = 1$ (depressed).

As shown in Figure 2, our E2-LPS operates as follows: The screening process identifies risky posts based on templates from the Beck Depression Inventory II (BDI-II) (Beck et al., 1996), a widely recognized clinical tool in psychology, covering both direct and indirect depressive indicators (e.g., pessimism, loss of appetite). For further details, please refer to Table 4 in the Appendix. We represent the results of the screening as a mask $\mathbf{m} = \{m_i\}_{i=1}^n$, where $m_i = 1$ indicates post $p_i$ is selected as risky.

### 3.1 Screening Process

During screening, a template set $S = \{s_1, \ldots, s_k\}$ from psychological scales is used, where each $s_i$ is a symptom template. Posts and templates are encoded using Sentence-BERT (Reimers, 2019) into embeddings $\mathbf{P} = \{\mathbf{p_1}, \ldots, \mathbf{p_n}\}$ and $\mathbf{S} = \{\mathbf{s_1}, \ldots, \mathbf{s_k}\}$. The risk score $r_i$ for each post $p_i$ is calculated by finding its maximum cosine similarity with any symptom template in $S$:

$$r_i = \max_{j=1}^{k} \frac{\mathbf{p_i}^T \mathbf{s_j}}{\|\mathbf{p_i}\| \|\mathbf{s_j}\|} \tag{1}$$

We compute the maximum cosine similarity between each post and all symptom templates. The scores for all posts are $\mathbf{r} = \{r_1, \ldots, r_n\}$. Posts corresponding to the top $K\%$ highest risk scores $r_i$ are selected as representative risky posts. The screening mask $\mathbf{m} = \{m_i\}_{i=1}^n$ is defined as $m_i = 1$ if $r_i$ is in the top $K\%$, and 0 otherwise. A straight-through estimator is employed to enable gradient flow through the discrete selection process represented by $\mathbf{m}$, facilitating end-to-end optimization. Additionally, the attention mechanism in the detection process is modified to facilitate gradient propagation back to the screening process, ensuring differentiability throughout the E2-LPS model.

### 3.2 Ensuring Differentiability

The non-differentiable Top K operator hinders gradient flow to the screening parameters. To address this, we employ a straight-through estimator, attaching gradients from risk scores $r_i$ to their corresponding mask entries $m_i$. Given the set of risk scores $\mathbf{r} = \{r_1, r_2, \ldots, r_n\}$, we compute the threshold $\tau$ as the $(100 - K)\%$ quantile of these scores. A hard selection mask $\mathbf{m}'$ is then generated using the indicator function $\mathbf{m}' = \mathbb{I}_{\{r \geq \tau\}}$. To achieve differentiability, we introduce a mask $\mathbf{m}$:

$$\mathbf{m} = \mathbf{r} + (\mathbf{m}' - \bar{\mathbf{r}}) \tag{2}$$

where $\bar{\mathbf{r}} = \text{detach}(\mathbf{r})$, representing the stop-gradient version of $\mathbf{r}$. We facilitate gradient flow by incorporating the linear term $\mathbf{r}$.

**Forward Pass Equivalence:** During the forward computation, since $\bar{\mathbf{r}}$ is merely a stop-gradient version of $\mathbf{r}$, we can express:

$$\mathbf{m} = \mathbf{r} + (\mathbf{m}' - \bar{\mathbf{r}}) = \mathbf{m}' \tag{3}$$

This indicates that in the forward pass, the hard selection mask $\mathbf{m}'$ is identical to the screening mask $\mathbf{m}$. Thus, the model's behavior remains unchanged during the forward propagation.

**Backward Pass Differentiability:** In the backward pass, since both $\mathbf{m}'$ and $\bar{\mathbf{r}}$ are treated as constants (i.e., without gradients), we can differentiate with respect to $\mathbf{r}$:

$$\frac{\partial \mathbf{m}}{\partial \mathbf{r}} = \frac{\partial}{\partial \mathbf{r}} [\mathbf{r} + \text{constant}] = 1 \tag{4}$$

Despite the non-differentiability of $\mathbf{m}' = \mathbb{I}_{\{r \geq \tau\}}$, the construction of $\mathbf{m}$ enables gradient propagation back to $\mathbf{r}$, making the process differentiable. It is evident that as long as the gradient can be propagated to $\mathbf{m}$, we can update $\mathbf{r}$ based on the gradient of $\mathbf{m}$, thereby adjusting the screening results.

### 3.3 Detection Process

In the detection process, E2-LPS focuses on identifying users with depression based on the screening results. However, it is essential to ensure that $\mathbf{m}$ receives updated gradients during this phase to guide the updating of the screening results. We utilize BERT to encode all posts made by the user and employ a Transformer encoder to model the interactions among various posts. The encoding equation is presented as follows:

$$\hat{\mathbf{p}}_\mathbf{i} = BERT_{[CLS]}(p_i) \tag{5}$$

Given the $n$ representations of a user's posts $\hat{\mathbf{p}}_\mathbf{1}, \hat{\mathbf{p}}_\mathbf{2}, \ldots, \hat{\mathbf{p}}_\mathbf{n}$, the user encoder models the relationships among these representations to generate updated contextual representations $\mathbf{p}'_\mathbf{1}, \mathbf{p}'_\mathbf{2}, \ldots, \mathbf{p}'_\mathbf{n}$ for each post, and further aggregate these embeddings into a unified user representation $\mathbf{u}$. We use a Transformer to capture inter-post dependencies via self-attention. The updated post representations are further passed to an attention layer, which produces the final detection results. To update $\mathbf{m}$ during training, we modify the masking mechanism for all attention layers as follows:

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^\mathbf{T}}{\sqrt{d}}, \quad \hat{A}_{i,j} = \frac{\exp(A_{i,j}) m_j}{\sum_{k=1}^{n} \exp(A_{i,k}) m_k} \tag{6}$$

In this modification, the self-attention mechanism ensures that during training, posts where $m_i = 0$ do not participate in subsequent computations; only posts with $m_i = 1$ are considered. The detection model requires access to all posts during training to ensure that all posts from the screening process receive gradient feedback from the detection process. However, during inference, we can discard the posts where $m_i = 0$ after the screening process, which does not increase the model's inference time.

We use the transformer encoder using a modified attention mechanism, allowing screened posts to interact with one another in a differentiable manner. We also apply this approach to enhance the general attention mechanism used for aggregating all selected information. Consequently, the user encoder can be represented as:

$$\mathbf{p}'_\mathbf{1}, \mathbf{p}'_\mathbf{2}, \ldots, \mathbf{p}'_\mathbf{n} = \text{Transformer}(\hat{\mathbf{p}}_\mathbf{1}, \hat{\mathbf{p}}_\mathbf{2}, \ldots, \hat{\mathbf{p}}_\mathbf{n}) \tag{7}$$

$$\alpha_l = \frac{\exp(\mathbf{W}'\mathbf{p}'_\mathbf{l} + b) m_l}{\sum_{i=1}^{n} \exp(\mathbf{W}'\mathbf{p}'_\mathbf{i} + b) m_i} \tag{8}$$

$$\mathbf{u} = \sum_{l=1}^{n} \alpha_l \mathbf{p}'_\mathbf{l} \tag{9}$$

The probability of depression based on the user representation is computed as:

$$\hat{y} = \text{sigmoid}(\text{FFN}(\mathbf{u})) \tag{10}$$

| | eRisk2017 | | eRisk2018 | | SMHD | |
|---|---|---|---|---|---|---|
| | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| No. of users (Depression) | 83 | 52 | 135 | 79 | 2,482 | 327 |
| No. of users (Control) | 403 | 349 | 752 | 741 | 4,964 | 2,054 |
| No. of posts (Depression) | 30,851 | 18,706 | 49,557 | 40,665 | 242,033 | 30,653 |
| No. of posts (Control) | 264,172 | 217,665 | 481,837 | 504,523 | 522,008 | 210,650 |
| Avg. posts per user (Depression) | 370.3 | 359.3 | 367.4 | 514.0 | 97.5 | 93.7 |
| Avg. posts per user (Control) | 656.0 | 623.0 | 641.8 | 680.2 | 105.2 | 102.6 |

Table 1: Dataset statistics for eRisk2017, eRisk2018, and SMHD. The table presents the number of users categorized by depression and control groups, the total number of posts for each group, and the average number of posts per user in both categories.

where $\hat{y} \in (0, 1)$ denotes the probability of depression based on the screened risky posts identified through the screening process. We utilize the cross-entropy loss function for end-to-end updates of the E2-LPS model.

## 4 Experiments

### 4.1 Datasets

In our experiments, we use three datasets: SMHD (Cohan et al., 2018), eRisk2017 (Losada and Crestani, 2016), and eRisk2018 (Losada et al., 2018). SMHD, designed for multi-disorder detection, allowed us to select depressed and control users. It contains 7,446 training users and 2,381 testing users. eRisk2017 includes 486 training users and 401 testing users. eRisk2018 comprises 887 training users and 820 testing users. These datasets identify depressed users through phrases like "I have been diagnosed with depression", while control users interact with depression-related content without a diagnosis. As can be seen, all datasets contain a large number of user posts, including substantial noisy signal. To avoid direct self-report leakage and prevent learning indirect depression signals, these datasets filter identification anchor points. Detailed dataset statistics are in Table 1.

### 4.2 Experimental Setups

We use `all-MiniLM-L6-v2` as the screening model and `bert-base-uncased` as the detection base model. The screening threshold is set at 12.5% (K%) per dataset. Batch size is 1; learning rates are $1 \times 10^{-5}$ for `BERT` and $2 \times 10^{-5}$ for other components. The user encoder is a single-layer `TransformerEncoder` with eight heads. We optimize with AdamW and train for 10 epochs. Training uses `PyTorch 2.2` on `Nvidia A100 40GB` GPUs. To mitigate randomness, we run each model three times with different seeds (Zhang et al., 2022; Liu et al., 2024).

### 4.3 Competing Methods

We compare our approach with several existing methods that researchers previously employed.

- **BERT (Clus+Abs)** (Zogan et al., 2021): This method employs Sentence-BERT (Reimers, 2019) for post embeddings, K-means clustering for representative post selection, and BART (Lewis et al., 2020) for abstractive summarization (Clus+Abs). The aim of this approach is to capture the content of the posting history through summarization.
- **HAN-BERT** (Zhang et al., 2022): This approach utilizes a Hierarchical Attention Network (HAN) with BERT as the post encoder. Its variant, **HAN-BERT (Psych)**, further employs a static scale-based screening method, filtering data by similarity based on a predefined scale.
- **COMMA** (Gui et al., 2019): This work's cooperative multi-agent model uses reinforcement learning to select textual and visual indicators for depression detection. We use its RL-driven text screening component as a baseline.
- **DeCapsNet** (Liu et al., 2024): Similar to HAN-BERT (Psych), this method screens data before training using a predefined scale based on similarity and introduces contrastive learning via symptom capsule extraction.
- **GPT-4.1** (Achiam et al., 2023): This explores the analytical capabilities of advanced LLMs in a 2-shot setting with balanced sampling

| Train | Method | Test: eRisk2017 | | Test: eRisk2018 | | Test: SMHD | |
|---|---|---|---|---|---|---|---|
| | | F1 | Recall | F1 | Recall | F1 | Recall |
| **eRisk2017** | COMMA | 0.552 | 0.552 | 0.500 | 0.519 | 0.509 | 0.842 |
| | Bert(Clus+Abs) | 0.531 | 0.500 | 0.468 | 0.468 | 0.573 | 0.502 |
| | HAN-BERT | 0.626 | 0.685 | 0.570 | 0.570 | 0.677 | 0.781 |
| | HAN-BERT(Psych) | 0.692 | 0.692 | 0.659 | 0.700 | 0.687 | 0.599 |
| | DeCapsNet | 0.699 | 0.705 | 0.614 | 0.646 | **0.748** | 0.700 |
| | GPT-4.1(Psych) | 0.556 | 0.769 | 0.484 | 0.785 | 0.706 | 0.714 |
| | GPT-4.1(E2-LPS) | 0.583 | 0.808 | 0.527 | 0.873 | 0.732 | 0.746 |
| | **E2-LPS(Ours)** | **0.714** | 0.821 | **0.688** | 0.706 | **0.748** | 0.894 |
| **eRisk2018** | COMMA | — | — | 0.511 | 0.443 | 0.550 | 0.775 |
| | Bert(Clus+Abs) | — | — | 0.515 | 0.557 | 0.602 | 0.560 |
| | HAN-BERT | — | — | 0.622 | 0.653 | 0.639 | 0.886 |
| | HAN-BERT(Psych) | — | — | 0.654 | 0.646 | 0.656 | 0.505 |
| | DeCapsNet | — | — | 0.641 | 0.747 | 0.753 | 0.731 |
| | GPT-4.1(Psych) | — | — | 0.467 | 0.798 | 0.707 | 0.722 |
| | GPT-4.1(E2-LPS) | — | — | 0.510 | 0.835 | 0.721 | 0.737 |
| | **E2-LPS(Ours)** | — | — | **0.683** | 0.727 | **0.777** | 0.907 |
| **SMHD** | COMMA | 0.532 | 0.500 | 0.478 | 0.443 | 0.743 | 0.723 |
| | Bert(Clus+Abs) | 0.520 | 0.635 | 0.438 | 0.608 | 0.663 | 0.685 |
| | HAN-BERT | 0.599 | 0.506 | 0.512 | 0.383 | 0.734 | 0.644 |
| | HAN-BERT(Psych) | 0.633 | 0.731 | 0.570 | 0.747 | 0.778 | 0.768 |
| | DeCapsNet | 0.605 | 0.705 | 0.527 | 0.747 | 0.779 | 0.795 |
| | GPT-4.1(Psych) | 0.566 | 0.827 | 0.458 | 0.835 | 0.693 | 0.741 |
| | GPT-4.1(E2-LPS) | 0.577 | 0.827 | 0.491 | 0.848 | 0.713 | 0.762 |
| | **E2-LPS(Ours)** | **0.639** | 0.697 | **0.611** | 0.566 | **0.797** | 0.902 |

Table 2: Main results of our experiments. The training set for eRisk2018 includes some user from eRisk2017; hence, model trained on eRisk2018 is not tested for eRisk2017. GPT-4.1(Psych) refers to results utilizing the same screening results as HAN-BERT(Psych). Conversely, GPT-4.1(E2-LPS) denotes results obtained using screening results consistent with our E2-LPS.

of depressed and healthy controls from the training set.

All implementations adhere to prior research methodologies. We evaluate both within-dataset and cross-dataset performance across three datasets. Additionally, we showcase the capabilities of advanced LLMs in mental health analysis using GPT-4.1. Concurrently, we demonstrate the challenging nature of this task.

## 5 Results

### 5.1 Main Result

We evaluate performance using two scenarios. The **within-dataset** scenario tests on the original dataset, while the **cross-dataset** scenario uses external datasets to specifically assess generalization capabilities. We aim to demonstrate that E2-LPS not only fits well on the specific dataset due to our training, but also exhibits generalization ability.

**Comparison in Within-Dataset Setting:** As demonstrated in Table 2, our method achieves better performance than other baseline models in the within-dataset scenario, particularly on the eRisk2018 and SMHD datasets. On the eRisk2017 dataset, our approach achieves comparable performance to the strongest baseline.

It is noteworthy that methods without a screening process perform worse than those incorporating screening, underscoring the importance of this process in user-level depression detection. Additionally, GPT-4.1(Psych)'s lower performance compared to GPT-4.1(E2-LPS) is interesting because only the input content differed, indicating E2-LPS screening more accurately identifies depressed users and improves performance. Simultaneously, COMMA, as a reinforcement learning method, performs well even when the average number of user posts is low. We believe this is due to its sampling process, which updates based on sampled

posts each time, and the absence of a psychiatric scale as an initial basis. Despite employing the most complex detection model, DeCapsNet's performance is limited by its screening results. Thus, E2-LPS is crucial for better depression detection. Therefore, our study highlights the important role of E2-LPS in enhancing the overall performance of depression detection methods.

**Comparison in Cross-Dataset Setting:** We test the trained models across different datasets. Table 2 reveals that our method outperforms baseline models in cross-dataset evaluations. This indicates that our gains do not solely originate from overfitting to the training dataset; instead, they reflect enhanced generalization capabilities, suggesting strong potential for practical deployment in diverse depression detection tasks.

We believe that the screening process in E2-LPS does not merely overfit to a specific training set; it effectively extracts critical information that contributes to improved model performance.

As shown in Table 5 in the Appendix B, the results of the one-tailed Welch's t-test indicate a statistically significant improvement in the performance of our model over the strong baseline ($p < 0.05$). Concurrently, we note a distinct variation in the recall scores. The data imbalance reflects a real-world scenario in which the recall score is a more critical metric for practical applications, especially when F1 scores are similar. This emphasis on recall is because it better reflects the model's ability to identify signals of depression from a large volume of data. Achieving a high recall score is challenging for this highly imbalanced task, given that the real-world prevalence of depression is substantially lower than that of non-depression. Therefore, the F1 score, in conjunction with the significance test, clearly demonstrates the superiority of our proposed method.

| Method | eRisk2017 | eRisk2018 | SMHD |
|---|---|---|---|
| | F1 | F1 | F1 |
| **E2-LPS** | **0.714** | **0.683** | **0.797** |
| w/o P | 0.640 | 0.634 | 0.693 |
| w/o LPS | 0.652 | 0.631 | 0.751 |
| w/o E2 | 0.657 | 0.640 | 0.748 |

Table 3: Results of the ablation analysis.

## 6 Ablation Analysis

We examine the effects of various modules within the E2-LPS model. We also examine the impact of jointly training the screening and detection processes and the effect of incorporating a psychiatric scale. As shown in Table 3, we evaluate three variants: E2-LPS w/o P, E2-LPS w/o LPS, and E2-LPS w/o E2.

- **E2-LPS w/o P**: We replace the psychiatric scale with a set of learnable vectors, enabling the model to learn independently of the psychological scale.
- **E2-LPS w/o LPS**: We freeze the screening process entirely during training, which fixes the parameters of the screening process.
- **E2-LPS w/o E2**:We train the screening process separately on a post-level labeled depression dataset (Turcan and McKeown, 2019) and use its output for screening.

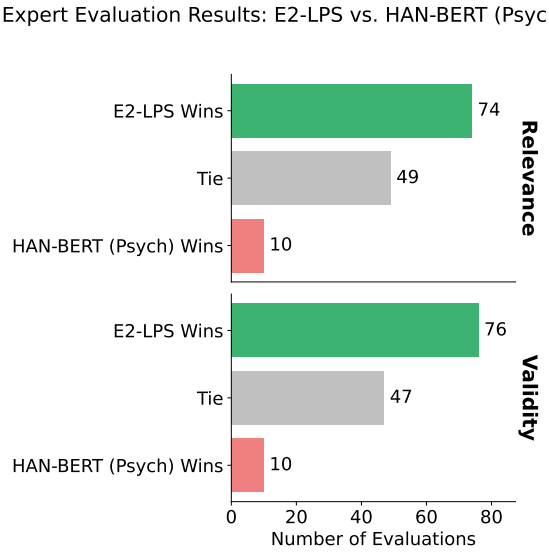


Figure 3: Expert evaluation of model screening effectiveness: E2-LPS vs. HAN-BERT (Psych).

Table 3 presents the ablation study results. The significant performance drop across the three variants underscores the critical role of each component within the E2-LPS framework.

## 7 Screening Analysis

This analysis examines two key aspects: the comparison of model screening results before and after training, and the impact of the percentage of posts used ($K\%$) on model performance.

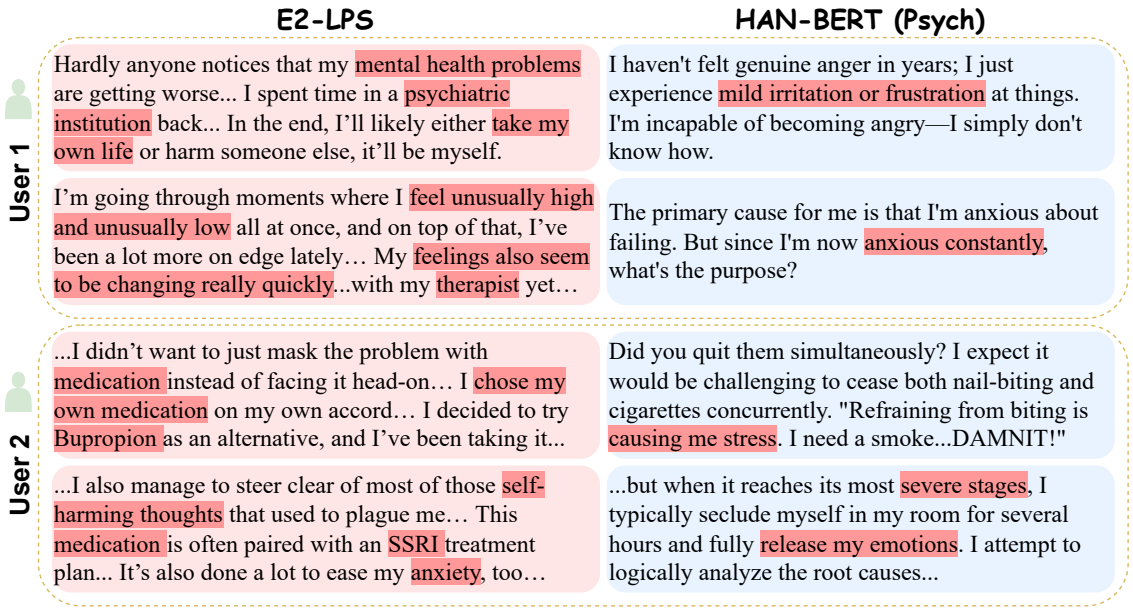


Figure 4: Two illustrative cases of screening results for depression users from the eRisk2018 dataset. The two most significant risky posts for each such user are presented, with relevant mental health-related sections highlighted in red. The content selected by E2-LPS notably includes more insights into users' psychological states, thereby demonstrating the effectiveness of our approach.

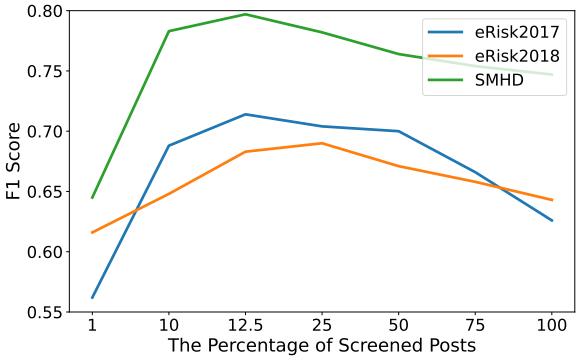


Figure 5: E2-LPS performance on percentage of posts used ($K\%$). Performance initially increases then decreases, indicating a complex relationship.

We conduct an expert evaluation where two experts compare the screening results of our method (E2-LPS) and an original HAN-BERT (Psych) method for posts in the eRisk2017 and eRisk2018 datasets. The comparison focuses on Depression Relevance (alignment with core depression symptoms) and Clinical Validity (practical value for assessment). The results, shown in Figure 3, indicate that E2-LPS enhances the model's ability to screen depression-related signals. Furthermore, the posts selected by E2-LPS hold greater potential to assist experts in accurately assessing users' mental states. We believe that providing experts with this information is even more significant than accuracy itself. Specific details are available in Appendix C.

Second, we investigate the effect of varying the risk post selection ratio ($K\%$) on model performance. As shown in Figure 5, initially, increasing $K\%$ improves depression detection by incorporating more relevant information. However, exceeding an optimal threshold leads to performance decline as the inclusion of excessive, potentially noisy posts outweighs the benefits. This underscores the importance of relevant input data for model performance and highlights the necessity of the screening process, thus confirming the influence of the screening process.Another key benefit of screening is its ability to highlight risky posts for professionals, facilitating human judgment.

## 8 Case Analysis

To further validate the efficacy of the E2-LPS model in screening users' mental health content compared to previous methods, this paper presents the screening results of different approaches. Specifically, two case studies of users with depression from the eRisk2018 dataset are showcased. **Due to the privacy protection limitations of the dataset, the original posts have been paraphrased to preserve the core meaning while ensuring that the original sentence structure or**

**specific, identifiable wording is not retained.** Figure 4 compares the top two posts screened by different methods (HAN-BERT (Psych) (Zhang et al., 2022)), which also uses `all-MiniLM-L6-v2` for screening. The results show that for both users, the outcomes from E2-LPS clearly surpass those of the Psych methods, better reflecting users' mental states. Our method captures not only more relevant mental health content but also captures more severe issues, including hospitalization, medication, and anxiety, highlighting the superiority of our approach.

## 9 Conclusion

In this paper, we introduce E2-LPS, an end-to-end learnable psychiatric scale-guided risky post-screening model for depression detection on social media. This model addresses the limitations of existing methods by jointly training a screening model, guided by psychiatric scales, alongside the detection model. The learnable screening process effectively extracts relevant mental health information from users' posting history, enhancing the detection performance. Our comprehensive experiments, conducted within and across datasets, demonstrate that E2-LPS outperforms several strong baseline methods. The results highlight the importance of integrating the screening and detection processes and the effectiveness of using psychiatric scales to guide the screening. Qualitative analysis further confirms that E2-LPS better captures users' mental states. [3]

## Limitations

We attempted to contribute to the detection and analysis of depression; however, we acknowledge that our article has potential limitations and may cause some harm.

- **False positive results:** Incorrectly identifying individuals as at-risk, which may lead to unnecessary interventions and psychological distress. We believe that screening results can help experts make informed decisions.
- **False negative results:** This occurs when the tool fails to identify individuals who genuinely require support. We emphasize that this tool is only a simple screening tool and cannot be used as evidence for a clinical diagnosis.
- **Potential stigmatization:** Users may be detected as depressed by the model without their knowledge, potentially violating their privacy and leading to discrimination in employment and internet-related aspects. We emphasize the correct use of research tools, which should not contribute to discrimination.

The findings from this study are preliminary. Therefore, the broader public, educators, and healthcare professionals should not rely on this or any similar computational model as a standalone diagnostic or screening tool. It should be viewed solely as an early-stage investigative instrument. Any future consideration for clinical utility will require extensive, independent, and rigorous validation, refinement, and ethical review to ensure its safety, efficacy, and alignment with established clinical best practices.

## Ethical Statement

This study involves a collaboration with a reputable academic medical center. The expert evaluators are board-certified psychiatrists, each possessing extensive clinical experience. The research on depression detection may raise certain ethical concerns. The data used in this study are acquired from publicly available datasets shared by other researchers. In order to protect individuals' privacy, all social media data undergoes strict anonymization procedures by the data providers before it is used. We comply with relevant ethical guidelines and legal regulations, ensuring that there is no risk of privacy violations during the research process. The E2-LPS is not intended as a diagnostic tool, but rather as a risk estimate for individual users that can then be used to support monitoring and evidence-based prevention and support for users.

## Acknowledgments

We are sincerely grateful to all the anonymous reviewers for their insightful and constructive comments, which were crucial for the improvement of our paper. We also wish to extend our gratitude to all the experts who participated in this study; their professional knowledge provided an important foundation for the successful progress of the research. This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project 2023ZD0121100, the National Natural Science Foundation of China (NSFC) via grant 62441614 and 62176078

[3] Our code will be released in: https://github.com/Nocturne-ZYZ/E2-LPS-EMNLP-2025

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report.

Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal Topic-Enriched Auxiliary Learning for Depression Detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1078–1089, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aaron T. Beck, Robert A. Steer, and Gregory K. Brown. 1996. *BDI-II, Beck Depression Inventory: Manual*. Psychological Corporation.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: A Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marta Coll-Florit, Salvador Climent, Marco Sanfilippo, and Eulàlia Hernández-Encuentra. 2021. Metaphors of Depression. Studying First Person Accounts of Life with Depression Published in Blogs. *Metaphor and Symbol*, 36(1):1–19.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–16, New York, NY, USA. Association for Computing Machinery.

J. Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, S. Marsella, D. Traum, A. Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of human and computer interviews. In *International Conference on Language Resources and Evaluation*.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19, pages 110–117, Honolulu, Hawaii, USA. AAAI Press.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):593–602.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22231–22239.

David E. Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language Use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer International Publishing.

David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 343–361, Cham. Springer International Publishing.

Flora Ma. 2021. Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5). In Danan Gu and Matthew E. Dupre, editors, *Encyclopedia of Gerontology and Population Aging*, pages 1414–1425. Springer International Publishing, Cham.

Anshu Malhotra and Rajni Jindal. 2022. Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, 130:109713.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the Latency of Depression Detection in Social Media. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Rafael Salas-Zárate, Giner Alor-Hernández, María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde, Maritza Bustos-López, and José Luis Sánchez-Cervantes. 2022. Detecting Depression Signs on Social Media: A Systematic Literature Review. *Healthcare*, 10(2):291.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *Preprint*, arXiv:1409.1556.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature Attention Network: Interpretable Depression Detection from Social Media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit Dataset for Stress Analysis in Social Media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Bichen Wang, Yixin Sun, Yuzhe Zi, Yanyan Zhao, and Bing Qin. 2024a. Scale-CoT: Integrating LLM with Psychiatric Scales for Analyzing Mental Health Issues on Social Media. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2651–2658.

Bichen Wang, Yuzhe Zi, Yanyan Zhao, Pengfei Deng, and Bing Qin. 2024b. 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.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022. Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression. arXiv.

Lina Zhou, Dongsong Zhang, Chris Yang, and Yu Wang. 2018. HARNESSING SOCIAL MEDIA FOR HEALTH INFORMATION MANAGEMENT. *Electronic Commerce Research and Applications*, 27:139–151.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 133–142.

## A Depression Templates

| No. | Statement |
|---|---|
| 1 | I feel depressed. |
| 2 | I am diagnosed with depression. |
| 3 | I am treating my depression. |
| 4 | I feel sad. |
| 5 | I am discouraged about my future. |
| 6 | I always fail. |
| 7 | I don't get pleasure from things. |
| 8 | I feel quite guilty. |
| 9 | I expected to be punished. |
| 10 | I am disappointed in myself. |
| 11 | I always criticize myself for my faults. |
| 12 | I have thoughts of killing myself. |
| 13 | I always cry. |
| 14 | I am hard to stay still. |
| 15 | It's hard to get interested in things. |
| 16 | I have trouble making decisions. |
| 17 | I feel worthless. |
| 18 | I don't have energy to do things. |
| 19 | I have changes in my sleeping pattern. |
| 20 | I am always irritable. |
| 21 | I have changes in my appetite. |
| 22 | I feel hard to concentrate on things. |
| 23 | I am too tired to do things. |
| 24 | I have lost my interest in sex. |

Table 4: Depression-related Statements

Here we provide the detailed templates in Table 4. Following prior work (Zhang et al., 2022), we employ the same combination of 3 direct depression descriptions and the 21 indirect symptoms derived from the Beck Depression Inventory-II (BDI-II) (Beck et al., 1996).

## B Significance Test

Here we supplement with the results of the significance test in Table 5, which further demonstrates the effectiveness of our method.

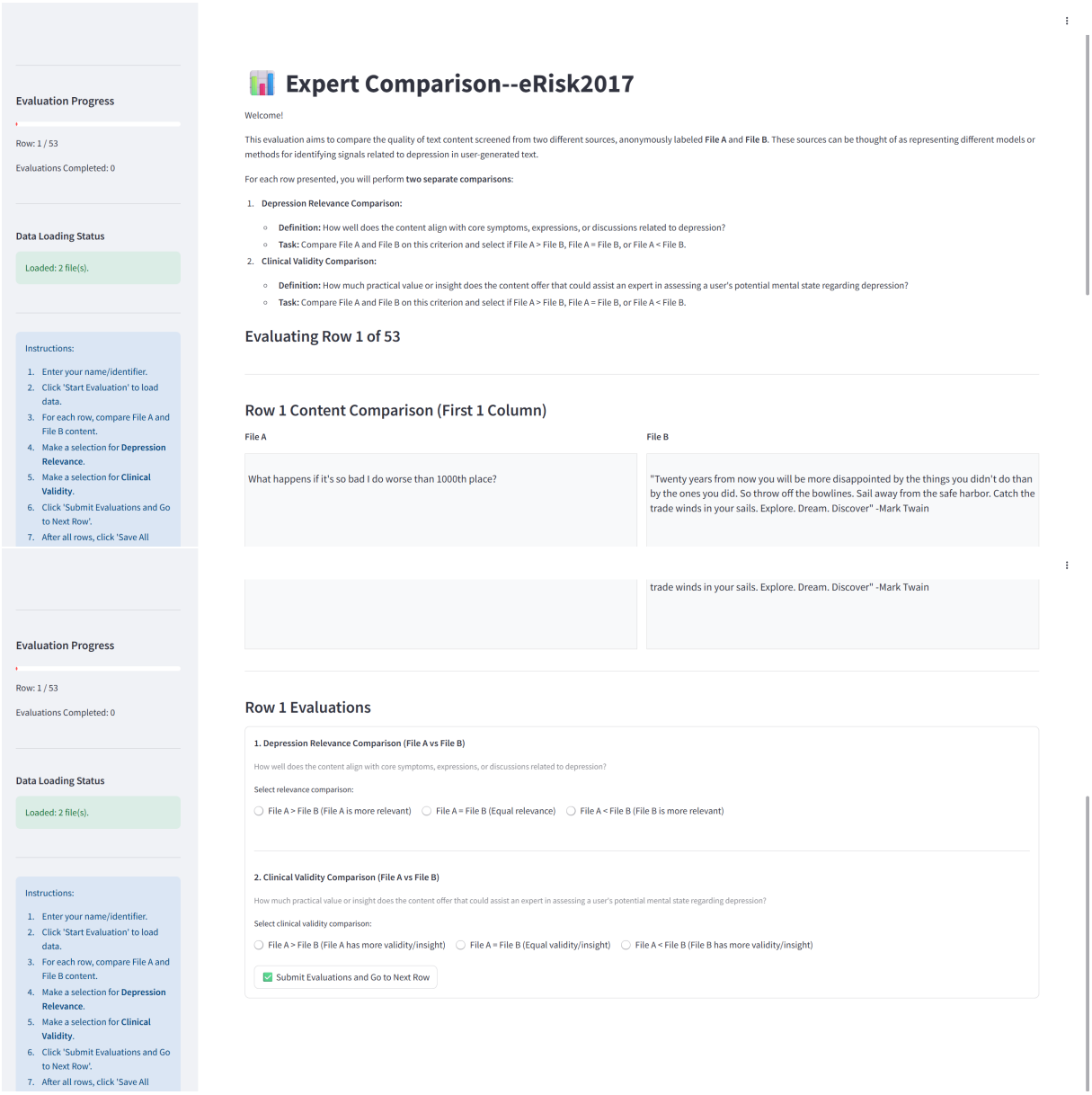


Figure 6: Web-based evaluation interface displaying paired posts from File A (left) and File B (right). Experts performed two independent comparative judgments per pair based on predefined criteria.

| Dataset | Baseline Model | P-Value |
|---|---|---|
| eRisk2017 | HAN-BERT | 0.0003 |
| | HAN-BERT(Psych) | 0.0214 |
| | DeCapsNet | 0.0348 |
| eRisk2018 | HAN-BERT | 0.0078 |
| | HAN-BERT(Psych) | 0.0095 |
| | DeCapsNet | 0.0016 |
| SMHD | HAN-BERT | 0.0009 |
| | HAN-BERT(Psych) | 0.0299 |
| | DeCapsNet | 0.0301 |

Table 5: Significance Test Results of Our Method against the Baseline Model on Various Datasets.

# C Expert Evaluation Scheme

## C.1 Overview and Purpose

To assess the relevance and clinical utility of depression-related content identified by different screening approaches, an expert evaluation was conducted. Two senior experts participate in this study. The evaluation aimed to compare content selected by our proposed model (E2-LPS) against content from a baseline model (Psych).

## C.2 Evaluation Interface and Procedure

The evaluation materials consist of 133 pairs of user posts. These posts were all sampled from de-

pressed users within the eRisk2017 and eRisk2018 test datasets. For each of the 133 selected depressed users, each pair includes the post identified by our proposed model as having the highest risk score for that user, and the post identified by the baseline model as having the highest risk score for the same user. To mitigate potential bias, the source of each post within a pair is anonymized and presented to the experts simply as "File A" and "File B".

A custom-developed web-based interface was used for the evaluation (see Figure 6). For each pair, posts from "File A" and "File B" were displayed side-by-side. Experts were instructed to perform two independent comparative judgments for each pair of posts based on the following criteria:

1. **Depression Relevance:** This criterion assessed how well the content aligned with core symptoms or expressions typically associated with depression. For this, experts chose one of the following options:
    - File A is more relevant than File B (A > B)
    - File A and File B are equally relevant (A = B)
    - File B is more relevant than File A (A < B)
2. **Clinical Validity:** This criterion evaluated the practical value or insight the content offered that could assist a clinical expert in assessing a user's potential mental state regarding depression. For this, experts chose one of the following options:
    - File A has more clinical validity/insight than File B (A > B)
    - File A and File B have equal clinical validity/insight (A = B)
    - File B has more clinical validity/insight than File A (A < B)

The experts proceeded through the evaluations item by item, submitting their judgments for both criteria before moving to the next pair.

### C.3 Inter-Rater Reliability

To quantify the level of agreement between the two experts, Cohen's Kappa ($\kappa$) coefficients were calculated independently for each evaluation criterion. The results are presented in Table 6. The Kappa value for Depression Relevance was 0.63 and for Clinical Validity was 0.74. According to the guidelines proposed by Landis and Koch (1977), both these values indicate Substantial agreement between the experts. These levels of agreement suggest a reliable basis for the evaluation outcomes.

| Evaluation Criterion | Cohen's Kappa ($\kappa$) |
|---|---|
| Depression Relevance | 0.63 |
| Clinical Validity | 0.74 |

Table 6: Cohen's Kappa ($\kappa$) Scores for Inter-Rater Reliability