HIT-SCIR@eRisk2025: Exploring the Potential of a Learnable Screening Model and Risk Post Buffer-Based Framework for Contextualized Early Prediction of Depression on Social Media

Notebook for the Task 2: Contextualized Early Detection of Depression Lab at CLEF 2025

Yuzhe Zi[1,†], Bichen Wang[1,†], Yanyan Zhao[1,*] and Bing Qin[1]

[1]*Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology*

Abstract

Task 2 of the eRisk lab at CLEF 2025 focuses on the contextualized early detection of depression using user posts from Reddit. The HIT-SCIR team participate in this task, submitting five runs based on different configurations of our proposed Learnable Screening Model and Risk Post Buffer-Based Framework. Our approach involves several key components: contextual data augmentation using Large Language Models (LLMs) to simulate social interactions and generate summaries for training data; a core end-to-end learnable risky post screening model guided by symptom descriptions from established psychiatric scales; and a depression risk detector utilizing MentalBERT for classification. **The official results on the test data demonstrate that our framework ranked first across several evaluation metrics, notably F1-score, ERDE50, Flatency, and various ranking-based measures.** This note describes the architecture, experimental setup, and performance analysis of our system, highlighting the value of integrating psychiatric knowledge into a learnable, context-aware model.

Keywords

Early Depression Detection, Social Media, Psychiatric Scale, Contextualized Detection

1. Introduction

According to the World Health Organization (WHO), depression affects approximately 3.8% of the global population[1]. In the United States, nearly 15% of adults experience at least one major depressive episode during their lifetime [1]. Early risk prediction represents an emerging research area with broad applications, such as identifying individuals at risk of mental disorders—a prominent societal concern. In depression detection, Early Risk Detection (ERD) is particularly crucial due to its predictive timeliness, as early warnings facilitate more timely intervention windows.

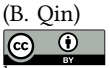
With the proliferation of the internet, social media platforms have become conventional avenues for individuals to openly express their thoughts and emotions [2]. Data from these platforms offer abundant resources for sentiment analysis and mental health inference [3]. Globally, considerable research focuses on leveraging social media for depression detection to mitigate the severe consequences of this condition [4, 5]. The CLEF eRisk 2025 Task 2 [6, 7] introduces a novel scenario for depression detection by incorporating complete conversational contexts. Unlike previous eRisk editions, which released only isolated posts from individual users, this year's task provides entire Reddit discussion threads involving the target user. This allows participating systems to access not only the target user's posts but also all interactions and their relational structures within the discussion threads. The task aims to simulate real-world scenarios where identifying depression necessitates the analysis of multi-party

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

yuzhezi@ir.hit.edu.cn (Y. Zi); bichenwang@ir.hit.edu.cn (B. Wang); yyzhao@ir.hit.edu.cn (Y. Zhao); qinb@ir.hit.edu.cn (B. Qin)



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

conversations. This edition features a dataset constructed from user posts on the Reddit platform. Our team, HIT-SCIR, participated in this task and achieved strong performance.

This study explores the effectiveness of a technique that combines an online detection algorithm based on a dynamic queue of at-risk posts [8] with an end-to-end learnable risk post screening scheme guided by psychological questionnaires, for the eRisk 2025 Task 2. Addressing the specifics of Task 2, which does not provide training data with user interactions, we employ LLMs to augment posts with potential contextual information to construct relevant training data.

Inspired by the implementation of Zhang et al. [8], our architecture comprises a **risky post screener** and a **depression risk detector**. During training, the screener calculates a risk score for each post based on its cosine similarity with descriptions from psychological scales. Subsequently, posts exhibiting higher risk scores are filtered for depression risk detection. The detector employs a Hierarchical Attention Network (HAN), utilizing BERT to acquire embedding representations for individual posts. It then models inter-post interactions using a Transformer and attention mechanisms, ultimately generating user features. Given the task's strong dependence on psychological knowledge, we leverage MentalBERT[9], pre-trained on psychology-related data, to generate post embeddings. Furthermore, the Straight-Through Estimator (STE) technique is employed to jointly train the screener and the detector. For final early detection testing, we utilize a dynamic risky post queue in conjunction with different alerting strategies for early depression detection. Through five-fold cross-validation, we evaluate the model's performance under various parameter settings. The top three performing models are selected for ensemble voting, and different early decision strategies are configured across five submission runs. The results indicate that our method outperforms other participating systems on most evaluation metrics.

The screener calculates a risk score for each post based on its cosine similarity with descriptions from psychological scales. Subsequently, posts exhibiting higher risk scores are filtered for depression risk detection. The detector utilizes MentalBERT to acquire embedding representations for individual posts and their associated interaction information. It then models inter-post interactions using a Transformer and attention mechanisms, ultimately generating user features. Furthermore, the Straight-Through Estimator (STE) technique is employed to jointly train the screener and the detector. The screener updates its results based on the detector's results. For final early detection testing, we use a dynamic risky post queue in conjunction with different alerting strategies for early depression detection. We evaluate the model's performance under various parameter settings through five-fold cross-validation. The top three performing models are selected for ensemble voting, and different early decision strategies are configured across five submission runs. The results indicate that our method outperforms other participating systems on most evaluation metrics.

2. Proposed Frameworks

The remainder of this paper is structured as follows: Section 2 details the technical framework, Section 3 reports and analyzes the experimental results, and Section 4 concludes the study and discusses future directions.

Our overall process begins with a **Contextual Data Augmentation** phase. The subsequent framework, as illustrated in Figure 1, comprises two core stages: **Psychiatric Scale-guided Risky Post Screening** and **Dynamic User-Level Early Risk Assessment Strategy**.

2.1. Contextual Data Augmentation

The training data provided in eRisk2025 Task 2 consists of isolated user posts that inherently lack interactive context (e.g., comments or replies). However, posts encountered in test scenarios typically include associated contextual information, which proves crucial for accurate understanding and prediction. To bridge this discrepancy between the context-scarce training data and potentially context-rich test environments, and to enable our model to effectively leverage contextual information during testing, we

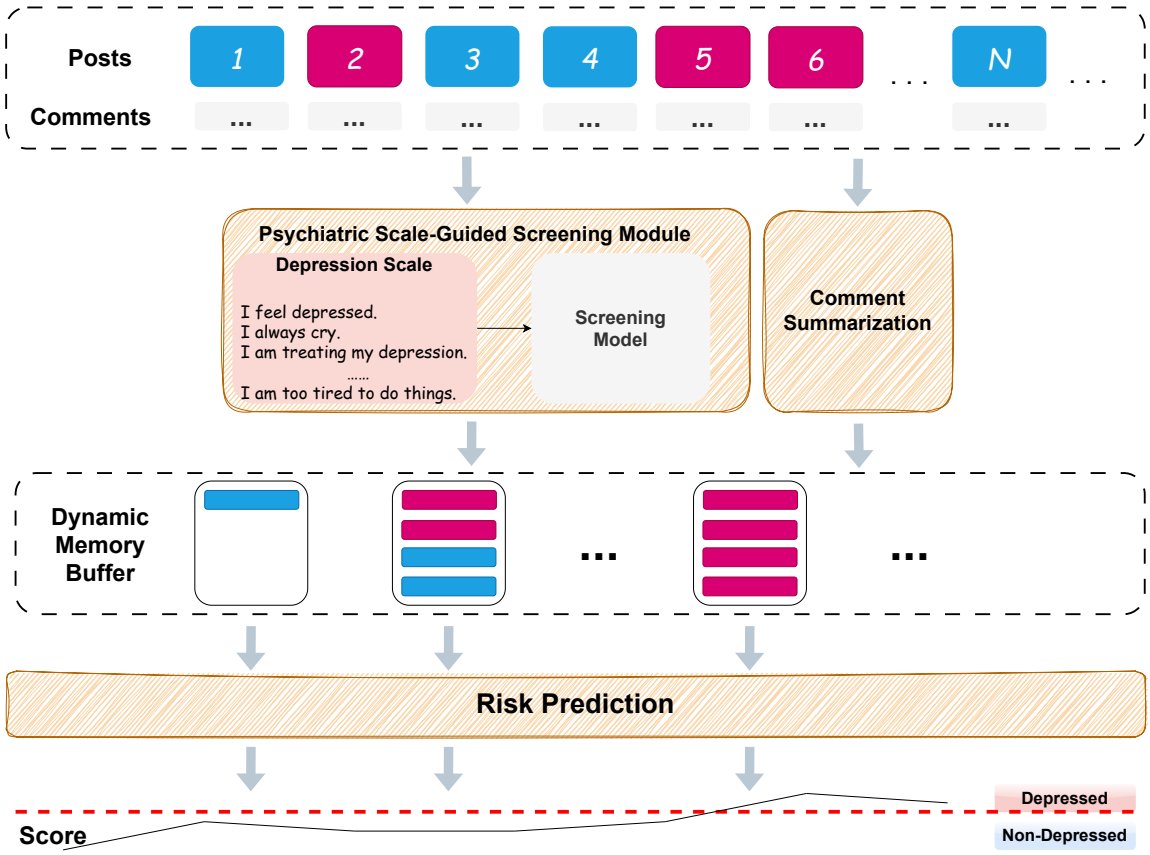


Figure 1: An overview of the two key components of our framework: Psychiatric Scale-guided Risky Post Screening and Dynamic User-Level Early Risk Assessment Strategy.

employ LLMs to generate simulated contextual information for the original user posts in the training data. This stage involves two specific steps:

Generation of Simulated Social Interactions: For each original post P_i in the training set, a pre-trained generative LLM, denoted as LLM_{gen} (e.g., we use the Phi-4 model [10] in this study), is utilized to generate a set of simulated comments $\{C_{i,1}, C_{i,2}, \ldots, C_{i,N_c}\}$. Here, N_c is a pre-defined number of generated comments. This process is guided by a specific instruction prompt, $\text{Prompt}_{\text{comment}}$, which directs the model to generate diverse and relevant comments based on the post content, outputting them in a structured format (e.g., JSON). This process can be represented as:

$$\{C_{i,j}\}_{j=1}^{N_c} = \text{LLM}_{\text{gen}}(P_i, \text{Prompt}_{\text{comment}}) \tag{1}$$

This step aims to supplement the original post with potential social reactions and discussion points, thereby providing a richer semantic context.

Summarization of Comment Content: Considering that multiple generated comments might contain redundant information or introduce unnecessary noise, we further summarize the comment set $\{C_{i,j}\}$ generated in the previous step to extract its core semantics. This step also employs an LLM, LLM_{sum} (in this study, we use the Phi-4 model), guided by a specific summarization instruction prompt, $\text{Prompt}_{\text{summary}}$. The generated summary is denoted as S_i.

$$S_i = \text{LLM}_{\text{sum}}(\{C_{i,j}\}_{j=1}^{N_c}, \text{Prompt}_{\text{summary}}) \tag{2}$$

Through this contextual augmentation process, each original post P_i is equipped with a highly condensed comment summary S_i generated by the LLM. Both P_i and S_i serve as inputs to the subsequent risk prediction model.

2.2. Psychiatric Scale-guided Risky Post Screening

We focus on how to extract key depression-related information from users' post histories. Our approach involves using psychological scales to screen for useful posts and then detect depression. To prevent cascading errors, we have implemented an end-to-end joint training approach for both the screening and detection stages. This model facilitates a seamless flow between identifying relevant posts and ultimately detecting depression.

2.2.1. Input Feature Representation

The model receives the user's original post P_i and its corresponding augmented context (comment summary S_i) as input. We employ a pre-trained text encoder, $\text{Encoder}_{\text{FE}}$, which is MentalBERT [9] — a BERT variant optimized for mental health texts — to independently convert these two text segments into fixed-dimensional dense vector representations:

$$\mathbf{e}_{P_i} = \text{Encoder}_{\text{FE}}(P_i) \tag{3}$$

$$\mathbf{e}_{S_i} = \text{Encoder}_{\text{FE}}(S_i) \tag{4}$$

Subsequently, these two embedding vectors are concatenated to form the comprehensive feature representation for the post, $\mathbf{x}_i = [\mathbf{e}_{P_i}; \mathbf{e}_{S_i}]$.

2.2.2. Psychiatric Scale-Guided Screening Module

This module aims to assess the importance of each post based on psychiatric scales and to screen for representative posts.

Construction Symptom Templates and Embeddings: For the screening process, we first define a scale item template set $Q = \{q_1, q_2, \ldots, q_M\}$. These entries q_k (symptom templates) are derived from widely used and validated psychological depression assessment scales [11], and are supplemented by three direct descriptive sentences for depression. All descriptions are detailed in Appendix A. These q_k represent clinically recognized criteria and manifestations of depression as defined by these scales. Each scale item template q_k is encoded into an embedding vector $\mathbf{e}_{q_k}$ using the text embedding model $\text{Encoder}_{\text{E5}}$, specifically multilingual-e5-large-instruct [12]. To enhance semantic alignment, we prepend a task-specific prompt specifically to these scale item descriptions before they are input to $\text{Encoder}_{\text{E5}}$. The prompt used is:

```
Given a description of depression symptoms, retrieve the posts from
user submissions that exhibit those symptoms.
```

This process for obtaining the scale item template embeddings $\mathbf{E}_Q = \{\mathbf{e}_{q_1}, \ldots, \mathbf{e}_{q_M}\}$ is represented for each template as:

$$\mathbf{e}_{q_k} = \text{Encoder}_{\text{E5}}(\text{Prompt} + q_k) \tag{5}$$

where 'Prompt' denotes the aforementioned task instruction, and $+$ signifies concatenation.

Post Risk Score Calculation: For each user post P_i, its textual content is directly encoded using the same $\text{Encoder}_{\text{E5}}$ to obtain the post embedding $\mathbf{e}'_{P_i}$. Crucially, no prompt is prepended to the post content for this embedding step:

$$\mathbf{e}'_{P_i} = \text{Encoder}_{\text{E5}}(P_i) \tag{6}$$

Then, we calculate the cosine similarity between this post embedding $\mathbf{e}'_{P_i}$ (from Eq. 6) and all scale item embeddings $\mathbf{e}_{q_k}$ (from Eq. 5) in the knowledge base. The maximum similarity value is taken as the importance score r_i of post P_i:

$$r_i = \max_{1 \le k \le M} \left(\frac{\mathbf{e}'_{P_i} \cdot \mathbf{e}_{q_k}}{\|\mathbf{e}'_{P_i}\| \|\mathbf{e}_{q_k}\|} \right) \tag{7}$$

The collection of importance scores for all posts is denoted as $\mathbf{r} = \{r_1, r_2, \ldots, r_L\}$.

Differentiable Dynamic Post Screening Mask Generation: Our goal is to select the top K' posts with the highest importance scores (where K' can be a proportion or a fixed number). Traditional Top-K' operations are non-differentiable. To enable end-to-end learning, we employ a Straight-Through Estimator (STE) method that allows for gradient propagation. First, based on the sequence of importance scores $[r_1, \ldots, r_L]$, a threshold τ is calculated (e.g., determined via quantile to identify the cut-off for the top K' scores). A preliminary hard selection mask $\mathbf{m}'$ is generated:

$$\mathbf{m}' = \mathbb{I}_{\{r \geq \tau\}} \tag{8}$$

where $\mathbb{I}$ is the indicator function. To achieve differentiability, we construct the final mask $\mathbf{m}$ as follows:

$$\mathbf{m} = \mathbf{r} + \mathbf{m}' - \text{detach}(\mathbf{r}) \tag{9}$$

where the detach$(\cdot)$ operation prevents gradients from flowing back through its arguments. This construction ensures that during the forward pass, the mask m_i behaves similarly to the hard selection m'_i, while during the backward pass, gradients can flow to the importance score $\mathbf{r}$ calculation, allowing parameters of the screening process to be optimized. This results in the mask sequence $\mathbf{m} = [m_1, m_2, \ldots, m_L]$.

2.2.3. Transformer Encoding and Final Classification

The mask mentioned above helps us filter out irrelevant information. We will now integrate this mask into subsequent operations in a differentiable form.

Mask-Guided Sequence Encoding: The user's sequence of comprehensive post features $[\mathbf{x}_1, \ldots, \mathbf{x}_L]$ and its corresponding differentiable dynamic feature selection mask sequence $[m_1, \ldots, m_L]$ are fed into one or more Transformer encoder layers (TransformerEncoderLayer). Within the self-attention mechanism, we modify the masking mechanism for all attention layers as follows:

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^{\mathbf{T}}}{\sqrt{D}}, \quad \hat{S}_{i,j} = \frac{\exp(S_{i,j})M_j}{\sum_{k=1}^{N} \exp(S_{i,k})M_k} \tag{10}$$

In this modification, where M_j represents the mask value m_j for the j-th post in the sequence being attended to, the self-attention mechanism ensures that during training, posts where $m_i = 0$ (referring to the mask of the i-th post in the overall user sequence) 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 updates to the screening process. However, during inference, we can discard the posts where $m_i = 0$, which does not increase the model's inference time. The output of the Transformer layers is:

$$[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_L] = \text{TransformerEncoderLayer}([\mathbf{x}_1, \ldots, \mathbf{x}_L], \text{attention\_mask} = [m_1, \ldots, m_L]) \tag{11}$$

Here, $\mathbf{h}_i$ is the context-aware representation of the post after Transformer encoding.

User-Level Representation Aggregation: To obtain a single vector representing the user's overall state, we aggregate the output sequence from the Transformer encoder $[\mathbf{h}_1, \ldots, \mathbf{h}_L]$. $\mathbf{h}_{\text{user}}$ is obtained through avg-pooling.

Risk Prediction: Finally, the aggregated user-level representation $\mathbf{h}_{\text{user}}$ is fed into a multi-layer perceptron (MLP) classifier, which outputs the probability $\hat{y}$ of the user having depression risk:

$$\hat{y} = \sigma(\text{MLP}(\mathbf{h}_{\text{user}})) \tag{12}$$

where σ is the Sigmoid activation function. The entire model is trained end-to-end by minimizing the binary cross-entropy loss (BCEWithLogitsLoss) between the predicted probabilities and the true labels. We jointly optimize the filtering of important posts and the final classification.

2.3. Dynamic User-Level Early Risk Assessment Strategy

To more authentically simulate the application scenarios of early risk detection, we adopt the following dynamic assessment process:

Dynamic Memory Buffer: For each user, we maintain a dynamic memory buffer of capacity K_{mem}. This buffer stores the K_{mem} posts from the user's recent history that are considered most relevant to depressive manifestations, based on their symptom relevance scores r_i. When a new post is published by the user, its r_i value is compared with those of the posts currently in the buffer. If the new post's relevance is higher and the buffer is full, the post with the lowest relevance is removed, and the new post is added. This ensures that the model always makes judgments based on the user's most relevant recent posts.

Sequential Prediction and Alert Mechanism: User posts are treated as a time series. At each time step (i.e., after a user publishes a new post and the memory buffer is potentially updated), the model (Section 2.2) performs a depression risk prediction based on the current set of posts in the memory buffer. If the model's predicted probability exceeds a pre-set decision threshold τ_{pred} for T_{con} consecutive times (e.g., $T_{con} = 2$), the system identifies the user as being at risk of depression and triggers an alert.

3. Experimental Evaluation

3.1. Datasets

The organizers provide task-specific corpora constructed from user posts on Reddit during designated time periods. The data is distributed in XML format, containing user IDs, timestamps, post titles, and textual content. For Task 2, the training corpus features binary classification labels distinguishing between depression cases and control groups. Notably, all classifiers for Task 2 were trained exclusively on this dataset without incorporating any external data sources. This approach ensures the evaluation reflects the models' performance under controlled conditions.

3.2. Experimental Setup

We submit the results of five runs, each based on a distinct decision-making approach for early detection. The core of all approaches is a risk prediction model implemented as a voting ensemble. This ensemble was constructed by selecting the three best-performing individual detection models from a candidate pool, a selection process guided by a five-fold cross-validation method on the training corpus. The five submitted runs (i.e., decision-making approaches) utilize this *same* pre-selected voting ensemble but differ in their operational parameters (τ_{pred}, T_{con}) as follows:

1. Run 0: τ_{pred}=0.5, T_{con}=1
2. Run 1: τ_{pred}=0.6, T_{con}=1
3. Run 2: τ_{pred}=0.7, T_{con}=1
4. Run 3: τ_{pred}=0.5, T_{con}=2
5. Run 4: τ_{pred}=0.5, T_{con}=3

The performance of the proposed frameworks on the training set is quantitatively assessed using precision (P), recall (R), and F1-score ($F1$) [13]. Additionally, the organizers evaluate the run results across multiple dimensions, including $ERDE_5$ [14], $ERDE_{50}$ [14], latency true positive rate ($latency_{TP}$) [6, 7], speed metric (speed) [6, 7], and latency-weighted F1-score ($F_{latency}$) [6, 7].

3.3. Analysis of Results

Table 1 presents the performance of the five HIT-SCIR test runs on decision metrics (Precision P / Recall R / $F1$-score) and temporal metrics ($ERDE_5$ / $ERDE_{50}$ / $latency_{TP}$ / speed / $F_{latency}$). Among all participating teams, HIT-SCIR-4 achieves the first rank in $F1$-score (0.85) and $ERDE_{50}$ (0.03), and second rank in $ERDE_5$ (0.06). Furthermore, HIT-SCIR-4 and HIT-SCIR-2 are tied for first place in the

Table 1
Decision-based evaluation for Task 2

Team	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$F_{latency}$
HIT-SCIR-0	0.72	**0.96**	0.82	**0.06**	**0.03**	**4.00**	0.99	0.81
HIT-SCIR-1	0.72	0.95	0.82	**0.06**	**0.03**	**4.00**	0.99	0.81
HIT-SCIR-2	0.74	0.94	0.83	**0.06**	**0.03**	**4.00**	0.99	**0.82**
HIT-SCIR-3	0.73	0.94	0.82	0.08	**0.03**	7.00	0.98	0.80
HIT-SCIR-4	**0.77**	0.94	**0.85**	0.09	**0.03**	8.00	**0.97**	**0.82**

Table 2
Ranking-based evaluation for Task 2

		Runs				
Writings	Metric	HIT-SCIR-0	HIT-SCIR-1	HIT-SCIR-2	HIT-SCIR-3	HIT-SCIR-4
1	P@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@100	**0.58**	**0.58**	**0.58**	**0.58**	**0.58**
100	P@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@100	**0.84**	**0.84**	**0.84**	**0.84**	0.83
500	P@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@100	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**
1000	P@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	$NDCG$@100	**0.90**	**0.90**	**0.90**	**0.90**	**0.90**

$F_{latency}$ metric (both 0.82). Regarding other metrics within the team: HIT-SCIR-0 exhibits the highest Recall (R) (0.96). HIT-SCIR-0, HIT-SCIR-1, and HIT-SCIR-2 perform identically and better than the other two submissions on $latency_{TP}$ (4.00) and speed (0.99); these three runs also outperform HIT-SCIR-3 (0.08) and HIT-SCIR-4 (0.09) on the $ERDE_5$ metric (0.06).

Ranking-based evaluation employs standard information retrieval metrics such as Precision at 10 (P@10) and Normalized Discounted Cumulative Gain ($NDCG$) to assess user risk levels sorted in descending order. Table 2 shows that the HIT-SCIR team's runs achieve first place in the vast majority of ranking metrics. Specifically: First, in the "1 writing" evaluation, all team runs rank first in P@10 (1.00) and $NDCG$@10 (1.00); however, the team does not achieve the top rank for the $NDCG$@100 metric (0.58). Second, when evaluating with one hundred posts ("100 writings"), all team runs rank first in P@10 (1.00) and $NDCG$@10 (1.00). For the $NDCG$@100 metric, the results of HIT-SCIR-0, HIT-SCIR-1, HIT-SCIR-2, and HIT-SCIR-3 (0.84) rank first, while HIT-SCIR-4 (0.83) performs slightly lower and does not achieve first place in this specific instance. Third, when evaluating with five hundred posts ("500 writings"), all team runs achieve first place in all three metrics: P@10 (1.00), $NDCG$@10 (1.00), and $NDCG$@100 (0.89). Finally, when evaluating with one thousand posts ("1000 writings"), all team runs also achieve first place in all three metrics: P@10 (1.00), $NDCG$@10 (1.00), and $NDCG$@100 (0.90).

4. Conclusion

The eRisk 2025 Task 2 underscores the complexities of early, contextualized depression detection from social media data. Our HIT-SCIR team proposes and evaluates a multi-stage framework centered around a learnable, psychiatric scale-guided screening model. This model is augmented by Large Language

Models (LLMs), which are employed for two key purposes: first, to perform contextual augmentation on the training data by simulating interactions, and second, to generate concise summaries of contextual information (e.g., comment threads) for both training and testing data. This summarization process offers the advantage of distilling relevant contextual signals while filtering out irrelevant information, thereby facilitating more focused and efficient processing by subsequent model components. Empirical analysis from our five submitted runs reveals the significant benefits of this integrated approach. The direct incorporation of knowledge derived from psychiatric scales into a differentiable screening mechanism, combined with the LLM-refined contextual information and the specialized representations from MentalBERT, allows our models to effectively identify and prioritize risk-indicative posts. This results in leading performance on key metrics such as $F1$-score, $ERDE_{50}$, and $F_{latency}$. While traditional methods might struggle with the nuanced and evolving nature of online discourse, our end-to-end learnable system demonstrates robust adaptability. Future work will focus on refining the contextual augmentation techniques, exploring more sophisticated modeling of multi-party conversational dynamics within threads, and further enhancing the screening module's sensitivity to subtle or emerging signs of depression. We also plan to investigate the framework's generalizability to other mental health conditions.

Declaration on Generative AI

During the preparation of this work, the authors use Gemini 2.5 Pro for grammar and spelling checks. After using this tool, the authors carefully review and edit the content as needed and take full responsibility for the publication's content.

References

[1] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, E. E. Walters, Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication, Archives of general psychiatry 62 (2005) 593–602.

[2] M. D. Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting Depression via Social Media, Proceedings of the International AAAI Conference on Web and Social Media 7 (2013) 128–137. doi:10.1609/icwsm.v7i1.14432.

[3] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 47–56. doi:10.1145/2464464.2464480.

[4] L. Zhou, D. Zhang, C. Yang, Y. Wang, HARNESSING SOCIAL MEDIA FOR HEALTH INFORMATION MANAGEMENT, Electronic Commerce Research and Applications 27 (2018) 139–151. doi:10.1016/j.elerap.2017.12.003.

[5] A. Malhotra, R. Jindal, Deep learning techniques for suicide and depression detection from online social media: A scoping review, Applied Soft Computing 130 (2022) 109713. doi:10.1016/j.asoc.2022.109713.

[6] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[8] Z. Zhang, S. Chen, M. Wu, K. Q. Zhu, Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression, arXiv, 2022. doi:10.48550/ARXIV.2205.09497.
[9] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190.
[10] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 Technical Report, 2024. doi:10.48550/arXiv.2412.08905. arXiv:2412.08905.
[11] A. T. Beck, R. A. Steer, G. K. Brown, BDI-II, Beck Depression Inventory: Manual, Psychological Corporation, 1996.
[12] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).
[13] T. Basu, S. Goldsworthy, G. V. Gkoutos, A Sentence Classification Framework to Identify Geometric Errors in Radiation Therapy from Relevant Literature, Information 12 (2021) 139. doi:10.3390/info12040139.
[14] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

A. Depression Templates

Table 3
Depression-related Statements

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.

Here we provide the detailed templates in Table 3. Following prior work [8], we employ the same combination of 3 direct depression descriptions and the 21 indirect symptoms derived from the Beck Depression Inventory-II (BDI-II) [11].