Contextualized Early Detection of Depression on Social Media using BDI-II Guided Screening and MentalBERT Ensemble

eRisk Task 2 - Contextualized Early Detection of Depression at CLEF 2026

Trung Xuan Bui[1], Tho Thi Ngoc Le[2,**]

Faculty of Information Technology, HUTECH University

Abstract

This paper describes the participation of the HUTECH-NLP team in Task 2 of eRisk 2026 - Contextualized Early Detection of Depression. This task aims to detect depression as early as possible from Reddit users by analyzing both the posts of each target user and the entire conversational threads in which the target user intervened. The conversational threads leverage the surrounding social context to make more reliable and timely decisions. To tackle this task, we propose a two-step pipeline. The first step filters depression posts through two successive filters: an E5 sentence encoder screens user posts with BDI-II symptoms to obtain candidate depression posts, and an ensemble of MentalBERT models classifies the candidates using context-enhanced posts. The ensemble is fine-tuned with a stratified group K-fold strategy partitioned by user to avoid data leakage, and trained with focal loss and class weighting to handle class imbalance. The second step detects depression users by aggregating post-level scores over the complete posting history into a user-level decision governed by a majority condition. We explore five configurations to observe different trade-offs with Precision, Recall, and Latency. On the dataset of 500 user threads in eRisk 2026 Task 2, we achieved our best F1 score of 0.74, with Precision of 0.77 and Recall of 0.70, while maintaining very low decision latency. The experimental results show that combining conversational context, symptom-based retrieval, and a leakage-free ensemble of domain-specific language models is effective for early depression detection.

Keywords

Early Depression Detection, Social Media, MentalBERT, BDI-II symptoms, Contextualized Detection

1. Introduction

Depression is among the most widespread mental health disorders in the world. According to the World Health Organization, as of 2025, approximately 332 million people live with depression, which is 5.7% of adults (4.6% of men and 6.9% of women) [1]. These numbers emphasize the severity of depression as a public health concern that extends across all regions and age groups. Despite the importance of the problem, access to mental care remains limited. In developed countries (e.g., the United States, the United Kingdom, and Australia), only about one third of patients affected by depression receive treatment, and the gap is considerably wider in least developed and developing countries [1]. This shortfall is due in part to social stigma, insufficient investment in mental health services, and a chronic shortage of trained professionals. As a consequence, a large proportion of cases remain undiagnosed and untreated. These remaining cases highlight a critical need for scalable, non-intrusive, and timely screening methods that can complement, rather than replace, traditional clinical assessment.

Clinical interviews are commonly used to assess an individual's depressive state at a single point in time. In contrast, social media generates continuous and naturally occurring streams of behavioral and linguistic signals. These data provide rich information for large-scale analysis of mental health–related signals. Recent advances in natural language processing and deep learning, and in particular Transformer-based language models, have demonstrated a strong capacity to capture the

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

** Corresponding author.

trung2441862032@hutech.edu.vn(buixuantrung1904@gmail.com) (T. X. Bui); ltn.tho@hutech.edu.vn (T. T. N. Le)

0000-0001-8068-8519 (T. T. N. Le)

subtle contextual cues associated with depression. Therefore, Transformer-based language models have motivated their application to early risk detection on the Internet.

1.1. Task 2 at eRisk 2026 - Contextualized Early Detection of Depression

The CLEF eRisk lab has been the reference evaluation forum for early risk prediction on the Internet since 2017. This forum provides annual shared tasks, datasets, and metrics which allow systems for the detection of mental health conditions to be compared under common conditions. Task 2 of the eRisk 2026 [2, 3] addresses the contextualized early detection of depression. Task 2 extends the challenge introduced in eRisk 2025 [4] by releasing the full Reddit discussion threads, in which a target user participates, rather than the isolated posts authored by that user.

This change enriches the information available to participating systems. The input for each target user comprises the posts written by the user, the comments and replies contributed by other participants, together with the relational structure that links the messages within multi-party conversations. The underlying motivation is that the clinical relevance of a message often becomes apparent only when it is interpreted alongside its surrounding context; a seemingly neutral reply, for instance, may reveal hopelessness when read as a response to criticism or a plea for help.

1.2. Contributions

This work describes the participation of the HUTECH-NLP team in Task 2 of eRisk 2026. The main contributions are summarized as follows:

- **A two-step detection pipeline.** The proposed system first filters depression posts through a screener guided by a psychiatric scale based on E5 and a fine-tuned MentalBERT ensemble, and then detects depression users by aggregating post-level scores into a decision governed by a majority condition, integrating clinical knowledge with domain-specific language modeling.
- **A majority condition for alerting.** An alert is triggered only when more than half of a user's posts exhibit depressive signals. This requirement substantially reduces false positives that would otherwise arise from a small number of isolated risk-bearing messages.
- **A full-history aggregation.** Eight complementary features are computed over the complete posting history of each user, without buffer limitations, so that long-term behavioural patterns are captured rather than only the most recent writings.
- **Diverse run strategies.** Five submission runs are configured to optimize different evaluation dimensions, ranging from early detection to balanced F1 and high-precision performance, allowing the trade-offs of the task to be explored systematically.

2. Related Work

2.1. Depression Detection from Social Media Text

Research on automatic depression detection from social media text has evolved considerably over the past decade [5]. Early systems relied on handcrafted features, such as sentiment lexicons, word frequency statistics, and linguistic markers, which were fed into traditional machine learning classifiers. These methods were gradually superseded by deep learning approaches, such as Convolutional Neural Networks and Long Short-Term Memory networks. These deep learning techniques learn discriminative representations directly from text and reduce the dependence on manual feature engineering.

The introduction of Transformer architectures mark a further turning point. Pre-trained language models such as BERT [6] and RoBERTa [7], together with domain-adapted variants such as MentalBERT [8], have achieved state-of-the-art results across a range of mental health text classification benchmarks. Their capacity to model contextual meaning allows them to capture subtle cues of psychological distress that lexicon-based methods tend to overlook.

2.2. MentalBERT and Domain-Specific Large Language Models

MentalBERT [8] is a Transformer language model initialized from BERT and further pre-trained on a large collection of mental health related posts gathered from Reddit communities. Through this domain adapted pre-training, the model acquires a vocabulary and a set of contextual representations that are aligned with the way users discuss emotional states online. As a result, MentalBERT consistently outperforms general purpose BERT on mental health classification tasks, including depression and suicidal ideation detection.

The success of MentalBERT reflects a broader trend toward specialized pre-training for mental health natural language processing. MentalBERT's performance shows that specialized in-domain training is an effective way to bridge the gap between general language understanding and the specialized terminology used in mental health.

2.3. Psychiatric-Scale-Guided Approaches

A complementary direction integrates clinical achievements directly into the detection pipeline. Zhang et al. [9] pioneered psychiatric-scale-guided risky post screening, in which the embedding of each post is compared, through cosine similarity, with textual descriptions of the items of clinical depression scales such as the BDI-II [10]. The most similar posts to a scale item are treated as risky and prioritized for further analysis, which both reduces the volume of text to process and provides an interpretable diagnostic basis grounded in clinical criteria.

The BDI-II [10] is a widely validated self-report inventory composed of 21 items, which cover the cognitive, affective, and somatic symptoms of depression, with scores grouped into minimal, mild, moderate, and severe severity levels. Screening with such a scale bridges the gap between data-driven natural language processing and the diagnostic standards used in clinical practice, and it forms the conceptual basis for the symptom-guided screening stage adopted in this work.

2.4. Ensemble Methods in Text Classification

Ensemble learning is a well established strategy for improving the robustness of text classifiers. Common approaches include hard voting, which selects the majority predicted label, soft voting, which averages the predicted probabilities of several models, and stacking, which trains a meta classifier on the outputs of base models. K-fold cross validation offers a natural source of diverse ensemble members, since each fold yields a model trained on a different partition of the data. Selecting only the strongest folds, rather than combining all of them, can further reduce the noise introduced by weaker models. Following this principle, the present work adopts soft voting over the three best folds, chosen according to validation F1, to form its depression classifier.

3. Methodology

Figure 1 illustrates our proposed two-step pipeline for detecting depression:

- Step 1: *Filtering depression posts.* User posts are classified into two categories: depression posts and normal posts. Section 3.1 details the process of filtering depression posts using pre-trained large language models.
- Step 2: *Detecting depression users.* Using depression posts and corresponding contexts of each user, we propose a formula to calculate a depression score. Then, the status of depression is decided by a chosen threshold. Section 3.2 details the calculation of the depression score.

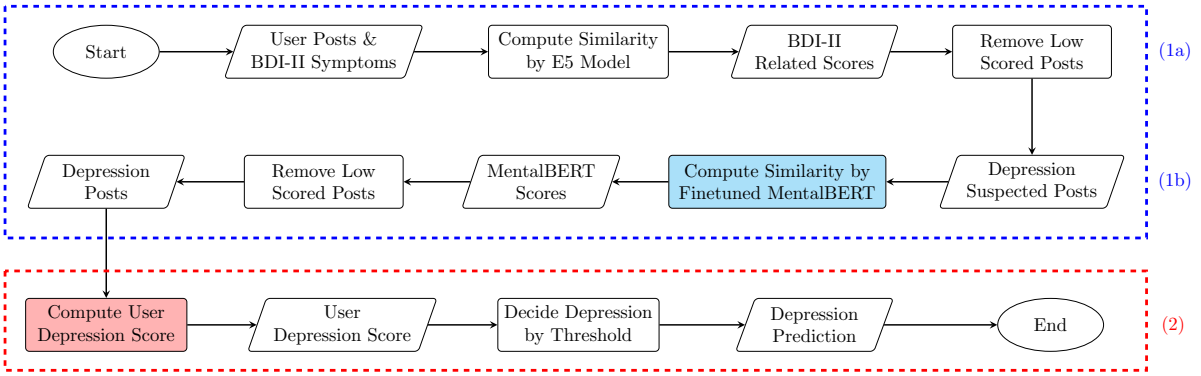

Figure 1: Overview of our proposed pipeline: Step (1) Filtering depression posts using two successive filters, and Step (2) Detecting depression users using our proposed formula.

3.1. Filtering Depression Posts

The first step reduces each user's writings to the depression posts, which are most likely to carry diagnostic signals. Two successive filters, both based on pre-trained large language models, are employed: the E5 model [11] and a fine-tuned MentalBERT [8].

- The first filter is based on BDI-II symptom similarity. Every post, paired with the BDI-II symptom descriptions, is passed to the E5 encoder to produce BDI-II-related scores for the post (Section 3.1.1). Then, posts whose scores fall below a threshold are removed, and the remaining posts form the set of depression-suspected posts. This step concentrates on clinically relevant content and discards posts that are unlikely to relate to mental states, such as casual conversation or topics unrelated to depression.
- The second filter is based on depression probability. The depression-suspected posts, as the results from the first filter, are scored by a fine-tuned MentalBERT model (Section 3.1.2). Posts with low MentalBERT scores are in turn discarded. At this step, the remaining writings are both related to the BDI-II symptoms and are confidently classified as depressive contribute to the user-level decision. To serve the second filter, MentalBERT has been adapted to the depression detection task and assigns each post a depression probability. The fine-tuning of MentalBERT is described in Section 3.1.3.

3.1.1. Compute Similarity by E5 Language Model

The symptom similarity filter is guided by the BDI-II inventory, which anchors the filter in an established diagnostic instrument rather than in surface keywords. It relies on a set of 24 short symptom descriptions derived from the BDI-II. Three of these are direct statements of depression, namely `I feel depressed.`, `I am diagnosed with depression.`, and `I am treating my depression.`, while the remaining 21 correspond to the individual BDI-II symptom items, covering manifestations such as sadness, discouragement about the future, feelings of failure, loss of pleasure, guilt, self-criticism, suicidal thoughts, changes in sleeping pattern, loss of energy, irritability, changes in appetite, difficulty concentrating, and fatigue. The complete list is provided in Appendix A.

Both the templates and the posts are embedded with the same E5-base encoder. Following the convention of the E5 model, each template is encoded with the `query:` prefix and each post with the `passage:` prefix. Sentence embeddings are obtained by mean pooling the last hidden states, weighted by the attention mask, and are subsequently normalised to unit length with L2 normalisation. The posts are encoded from their raw text so that the screening reflects the original wording of the user.

For each post i, a BDI-II-related score is defined as the maximum cosine similarity between its embedding and any of the 24 template embeddings:

$$r_i = \max_{1 \le k \le 24} \cos\left(\mathbf{e}_i^{\text{passage}}, \mathbf{e}_k^{\text{query}}\right). \tag{1}$$

Because the embeddings are L2 normalised, the cosine similarity in Eq. (1) reduces to a dot product. A high value of r_i indicates that the post is semantically close to at least one depressive symptom description.

The set of depression-suspected posts is determined by a dynamic threshold θ that adapts to the number of posts available for the user in the current round. When a user has more than three posts, the threshold is

$$\theta = \max\big(Q_{0.7}(\{r_i\}),\, 0.3\big), \tag{2}$$

that is, the larger value of the 70^{th} percentile of the observed scores and a minimum BDI-II threshold of 0.3. When a user has three or fewer posts, the percentile-based component is omitted and only the minimum BDI-II threshold $\theta = 0.3$ is applied, since a reliable quantile cannot be estimated from so few observations. All posts with $r_i \geq \theta$ are retained. If no post reaches the threshold, the single highest-scoring post is kept as a fallback, so that at least one post always proceeds to the MentalBERT filter.

3.1.2. Compute Similarity by Fine-tuned MentalBERT

Unlike the E5 filter works on raw post text, the MentalBERT classifier receives each depression-suspected post together with its conversational context. This context addresses a characteristic of the contextualized task, which is the relevance of a writing frequently depends on the conversation that surrounds it. Note that, following the eRisk terminology, we use the term *writing* to refer to any individual message authored by a user. Every writing of the target user is therefore formatted using one of two templates, according to whether contextual information is available.

When context is present, the input follows the template

`[CLS] [SOCIAL_CTX] {context} [SEP] [TARGET] {text} [SEP],`

and when no contextual information is associated with the writing, the template is reduced to

`[CLS] [NO_CTX] [SEP] [TARGET] {text} [SEP].`

The special markers `[SOCIAL_CTX]` and `[TARGET]` delimit the two information sources. The marker `[NO_CTX]` signals the absence of context so that the model can treat isolated writings consistently rather than confusing an empty context with a meaningful one. The context is drawn from the contributions of other participants in the same discussion thread: comments authored by users other than the target are collected, filtered to discard empty or removed content, and concatenated under per comment and overall length limits. For writings that correspond to a reply by the target user, the parent comment is prioritised as the context source, since it provides the most direct conversational antecedent; when the parent comment cannot be recovered, the system falls back to the general thread context.

3.1.3. Fine-tuning MentalBERT using eRisk 2025 Data

Training data and group K-fold split. The classifier is a MentalBERT model fine-tuned for post-level depression detection. The training corpus comprises 419,192 posts written by 909 users, of whom 807 are labelled as control and 102 as depressed, which corresponds to a prevalence of approximately 11.2%. Together with the variable number of posts per user, this imbalance requires a carefully designed data split and a class-weighted training objective. The data are partitioned with a group-aware variant of five-fold cross-validation. A `StratifiedKFold` split is first applied to a user-level table, with one row per user, stratified by the depression label so that the proportion of depressed users is preserved across folds. The resulting user-level partition is then mapped back to the post level, so that every post inherits the fold assignment of its author. Each validation fold contains approximately 181 to 182 users, of which around 161 are control and 20 to 21 are depressed.

A central property of this split is the prevention of data leakage: all posts written by a given user are assigned to exactly one fold. This requirement is verified through two assertions: each user appears in only one validation fold and no pair of folds shares any user. The motivation for splitting at the user level

rather than at the post level is based on an observation: a random split over posts would place writings of the same user in both the training and the validation set, and therefore the model is possibly memorise user-specific writing style and vocabulary instead of learning generalisable indicators of depression. Such leakage inflates validation metrics and yields an over-optimistic estimate of performance on unseen users. Because the deployment scenario of early detection involves users who has never been encountered during training the model, a user-level split is the appropriate evaluation protocol.

Training configuration. The base model is MentalBERT, a BERT model pre-trained on mental health corpora drawn from Reddit communities and related clinical text, loaded from a local copy so that training does not depend on any external service. To address the class imbalance, training minimises a class-weighted focal loss:

$$\mathcal{L}_{\text{focal}} = -\alpha_t \, (1 - p_t)^{\gamma} \, \log(p_t), \tag{3}$$

where p_t is the predicted probability of the ground truth class. The focusing parameter is set to $\gamma = 2.0$, which down-weights easy examples and concentrates the gradient on hard cases. The class weights α_t are obtained by an inverse-frequency rule that up-weights the depressed class to compensate for its roughly fivefold under-representation at the post level, and are then normalised so that the two weights sum to two.

The model is fine-tuned for two epochs with a learning rate of 1.5×10^{-5}. A per-device batch size of 64 combined with a gradient accumulation of four steps yields an effective batch size of 256. The learning rate follows a cosine decay schedule with a warmup ratio of 10%. Regularisation is provided by a weight decay of 0.05 and a label smoothing factor of 0.05. We adopt a partial layer-freezing strategy: the embedding layer and the first two encoder layers are frozen, while the remaining ten encoder layers, the pooler, and the classification head remain trainable. Freezing the lower layers reduces the risk of overfitting to the comparatively weak depressed-class signal, while retaining the domain-specific representations acquired during MentalBERT pre-training.

Ensemble construction. Once the five folds have been trained, they are ranked by their validation F1 score, and the three best folds, namely Fold 3, Fold 4, and Fold 1, are selected as the members of the ensemble, while the two weakest folds are excluded so that they do not dilute the aggregate performance. At inference time, each ensemble member independently estimates the probability that a post indicates depression, $P(\text{depressed} \mid \text{post})$, through a softmax over its output, and the final post level probability is computed as the arithmetic mean of the three predictions. A post is then counted as depressive when this probability satisfies $P(\text{dep}) \geq \delta$, with $\delta = 0.5$, which constitutes the second filter of Step 1 and feeds the user level aggregation described next.

3.2. Compute User Depression Score

The second step condenses the post-level predictions accumulated for each user into a single user-level depression score and then decides whether to issue an alert. All posts processed for a user are retained across rounds without any buffer limit, so that the score can draw on the complete posting history of user as evidence accumulates over the course of the interaction.

The user-level score combines eight complementary features, listed in Table 1, each capturing a different facet of the accumulated evidence. The features range from central tendency measures of the predicted depression probability, such as its mean over risky posts and over all posts, to extreme value summaries, such as the maximum and the 90$^{\text{th}}$ percentile, and to proportion-based indicators, such as the fraction of posts classified as depressive and the fraction of risky posts exceeding the post-level threshold. The mean BDI-II score links the aggregation back to the E5 filter, while the evidence feature is a stepped sufficiency indicator that increases with the number of available posts.

The raw user-level score is the weighted sum of the eight features, constrained to the unit interval:

$$S = \min\Big(1, \, \max\Big(0, \, \sum_{i=1}^{8} w_i f_i\Big)\Big), \tag{4}$$

Table 1
The eight features and their weights in the user-level depression score.

Feature	Weight	Description
avg_dep_risky	0.25	Mean depression probability across risky posts
max_dep	0.15	Maximum depression probability
p90_dep	0.10	90^{th} percentile of depression probability
dep_post_ratio	0.15	Fraction of posts with $P(\text{dep}) \geq 0.5$
high_ratio	0.10	Fraction of risky posts exceeding the threshold
avg_bdi	0.10	Mean BDI-II similarity score
avg_dep_all	0.10	Mean probability across all posts
evidence	0.05	Sufficiency indicator based on post count

where f_i and w_i denote the ith feature and its weight. The evidence feature itself takes one of five discrete levels, increasing from 0.2 for fewer than five posts to 1.0 for at least fifty posts, with intermediate levels at five, ten, and twenty posts.

To avoid premature alerts when limited evidence is available, the score is attenuated according to the total number of posts n observed for the user:

$$S_{\text{adj}} = \begin{cases} 0.70\,S & \text{if } n < 3, \\ 0.85\,S & \text{if } 3 \leq n < 5, \\ S & \text{otherwise.} \end{cases} \tag{5}$$

This penalty prevents confident alerts in the early rounds, when a user may have written only one or two posts, while leaving the score unchanged once a sufficient history has accumulated.

Decision with majority condition. A user is predicted as depressed when the adjusted score reaches the prediction threshold, $S_{\text{adj}} \geq \tau_{\text{pred}}$, for T_{con} consecutive rounds. The consecutive-round requirement guards against isolated spikes in the score: a round in which the threshold is not met resets the counter, so that an alert is issued only after a sustained pattern of risk. In the runs that enable it, a majority condition is imposed in addition: when the fraction of a user's posts classified as depressive ($P(\text{dep}) \geq 0.5$) does not exceed one half, the adjusted score is capped at $0.9\,\tau_{\text{pred}}$, which keeps it below the prediction threshold and thereby suppresses the alert. In effect, an alert is permitted only once a majority of the user's posts exhibit depressive signals, which substantially reduces false positives arising from a small number of isolated risk-bearing messages.

4. Experiments and Evaluations

4.1. Dataset

Regarding the fine-tuning of MentalBERT, the training corpus consists of 419,192 posts written by 909 users, of whom 807 are labelled as control and 102 as depressed. The resulting class distribution, 88.8% control against 11.2% depressed, reflects the severe imbalance that is characteristic of depression detection datasets and that motivates the class-weighted training objective described earlier.

4.2. Evaluation Metrics

Two families of metrics are used, following the official evaluation of the task. The decision based metrics assess the binary alerts issued for each subject and comprise precision P, recall R, and the F1 score computed with respect to the positive class, together with the early risk detection error at two penalty settings, ERDE_5 and ERDE_{50}, the median latency over true positives, $\text{latency}_{\text{TP}}$, the speed factor, and the latency weighted score F_{latency}. The ranking-based metrics assess the quality of the risk ranking

produced after each data release and comprise P@10, NDCG@10, and NDCG@100, reported after 1, 100, 250, and 500 writings.

4.3. Cross-Validation Results at the Post Level

Before the official submission, the classifier was assessed through the user disjoint five fold cross validation described in Section 3. Table 2 reports, for each fold, the F1 score, Recall, and Precision at the default decision threshold $\tau = 0.5$, the ROC-AUC, and the F1 obtained at the best per-fold threshold. Because the validation sets contain only users that are absent from the corresponding training set, these figures provide a realistic estimate of generalization to previously unseen users rather than an over-optimistic post-level score.

Table 2
Per fold cross validation results on user disjoint validation sets. The best value in each column is shown in bold.

Fold	F1	Recall	Precision	ROC-AUC	Best τ	Best F1
0	0.4518	0.7590	0.3216	0.8232	0.575	0.4976
1	0.4789	0.7754	0.3465	0.8309	0.625	0.5277
2	0.3427	0.5434	0.2503	0.6804	0.600	0.3611
3	**0.5315**	0.6600	**0.4448**	**0.8319**	0.550	**0.5544**
4	0.5205	0.7255	0.4058	0.8295	0.550	0.5395
Mean	0.4651	0.6927	0.3538	0.7992	—	0.4961
$\pm$ Std	0.0675	0.0830	0.0680	0.0590	—	0.0696

4.4. eRisk 2026 Online Submission Statistics

Table 3
Configurations for online submissions of HUTECH-NLP team.

Run	BDI-II Threshold	MentalBERT Threshold	User Depression Score τ_{pred}	Depression Post Rate	Consecutive Post
0	0.3	0.5	score > 0.5	rate $> 0\%$	1
1	0.3	0.5	score > 0.6	rate $> 0\%$	1
2	0.3	0.5	score > 0.7	rate $> 0\%$	1
3	0.3	0.5	score > 0.5	rate $> 50\%$	1
4	0.3	0.5	score > 0.5	rate $> 50\%$	2

In the official online evaluation, the HUTECH-NLP team submitted five runs and processed the complete set of 500 user threads released by the server. The configurations are shown in Table 3. All five submitted runs share the screening and classification pipeline and the two fixed thresholds, the minimum BDI-II threshold (0.3) applied in the E5 filter and the post-level depression threshold (0.5) applied to the MentalBERT probabilities, and differ only in the decision criterion applied to the user-level score. Runs 0, 1, and 2 vary the user depression score threshold τ_{pred} from 0.5 to 0.7 to trade recall for precision. Runs 3 and 4 additionally require more than half of the user's posts to be classified as depressive. Run 4 additionally requires two consecutive posts to be classified as depressive. The full interaction, measured from the first to the last response, was completed in two days and three hours, which places the system among the ten teams that covered the entire test collection.

4.4.1. Decision-Based Results

The decision-based results of the five runs are reported in Table 4. The best balanced behavior is obtained by run 2, which reaches an F1 of 0.74 with a precision of 0.77 and a recall of 0.70. The

Table 4
Task 2 decision-based results for the HUTECH-NLP runs. The best F1 and F_{latency} are in bold.

Run	P	R	F1	ERDE_5	ERDE_{50}	$\text{latency}_{\text{TP}}$	speed	F_{latency}
0	0.31	0.99	0.48	0.09	0.07	1.00	1.00	0.48
1	0.48	0.96	0.64	0.09	0.05	3.00	0.99	0.63
2	0.77	0.70	**0.74**	0.11	0.06	4.00	0.99	**0.73**
3	0.40	0.93	0.56	0.08	0.06	1.00	1.00	0.56
4	0.43	0.92	0.59	0.09	0.05	3.00	0.99	0.58

remaining runs favor recall over precision: runs 0, 1, 3, and 4 attain recall values between 0.92 and 0.99, but their precision falls between 0.31 and 0.48, which lowers their F1 accordingly. This pattern is consistent with the design of the decision stage, in which run 2 applies the highest score threshold and therefore issues fewer but more reliable alerts.

4.4.2. Ranking-Based Results

Table 5
Task 2 ranking-based results for the HUTECH-NLP runs. All five runs obtain identical values at each observation point.

	1 writing	100 writings	250 writings	500 writings
P@10	1.00	1.00	1.00	1.00
NDCG@10	1.00	1.00	1.00	1.00
NDCG@100	0.71	0.82	0.84	0.85

Table 5 reports the ranking-based results, measured after 1, 100, 250, and 500 writings. The five runs produce identical ranking scores at every observation point, which indicates that the user-level risk scores induce a stable ordering that is largely insensitive to the decision parameters that distinguish the runs. The quality of the top of the ranking is high throughout: P@10 and NDCG@10 reach the maximum value of 1.00 at all observation points, while NDCG@100 improves steadily from 0.71 after a single writing to 0.85 after 500 writings, reflecting the benefit of accumulating evidence over the course of the interaction.

These official results confirm the trends observed during cross-validation. The high-recall configurations are well suited to the early and exhaustive detection that the ranking metrics reward, whereas the higher threshold of run 2 yields the most balanced decision-based performance. Taken together, the combination of a symptom-guided screening stage, a leakage-free ensemble of MentalBERT models, and a full-history aggregation scheme produces accurate rankings and early, reliable alerts under the contextualized early detection setting.

4.5. Discussions

Impact of the Majority Condition

The majority condition has a measurable effect on the volume of alerts. A comparison of run 0 and run 3, which share the same prediction threshold $\tau = 0.5$ and consecutive requirement $T_{\text{con}} = 1$ and differ only in whether the condition is enabled. This behaviour is consistent with the intended role of the condition: it suppresses alerts for subjects whose aggregated score is driven by a small number of high-probability posts embedded in otherwise non-depressive content, a plausible signature of a false positive. The condition is therefore most appropriate when the cost of a false positive is high, for example when an alert triggers a downstream clinical review. For subjects with little posting activity, the condition may delay an alert; this effect is complemented by the evidence-based score penalty, which independently makes early alerts harder when the available history is short.

Limitations

Some limitations of the current work should be acknowledged.

- Firstly, the classifier scores each post in isolation, so that inter-post dependencies and the temporal trajectory of depressive expression are captured only at the aggregation stage rather than within the classifier itself.
- Secondly, the eight aggregation weights are set manually rather than learned; a held-out user-level subset could be used in future work to tune them.
- Thirdly, both the training and the test data originate from Reddit, so generalisation to other platforms such as forums or clinical chat remains to be validated.
- Lastly, the BDI-II templates and MentalBERT are English-centric, leaving room to improve multilingual robustness in the mental health domain.

5. Conclusions and Future Work

This work reports the HUTECH-NLP submission to Task 2 of eRisk 2026, the contextualized early detection of depression. The proposed system is a modular pipeline of two main steps: filtering depression posts and computing user-level depression scores. Step 1 combines each writing together with its conversational context, screens posts with a training-free module based on E5 embeddings of 24 BDI-II-derived statements, classifies the retained posts with a leakage-free ensemble of MentalBERT models, and aggregates post-level predictions into a subject-level decision governed by an evidence penalty and an optional majority condition. A user-level group cross-validation protocol ensured that all posts of a given user remained within a single fold, providing a realistic estimate of generalisation to unseen users, and the three best folds were combined through soft voting. Step 2 combines different features of the user's depression posts to compute the user-level depression score. On the official test set, the system reached a best F1 of 0.74 with very low decision latency, together with strong ranking quality.

Future work will address the limitations identified above. One direction concerns the temporal nature of depressive signals: since episodes are time-bounded in clinical practice, a time-decay weight on post probabilities together with a score cool-down, allowing the aggregated score to drift downward when no new depressive content appears, would support recovery monitoring as well as early detection. Further promising directions include modeling inter-post dependencies and temporal dynamics within the classifier, extending the symptom-guided screening to additional psychiatric scales, and evaluating the robustness of the approach across platforms and languages.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Claude in order to: Get suggested translation from Vietnamese into English, Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] World Health Organization, Depressive disorder (depression), https://www.who.int/news-room/fact-sheets/detail/depression, 2025. Accessed: 2026-05-24.

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and ADHD, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[3] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and ADHD (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.
[4] J. Parapar, A. Pérez, 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, CEUR Workshop Proceedings, 2025, pp. 1449–1473. URL: https://ceur-ws.org/Vol-4038/paper_117.pdf.
[5] W. B. Tahir, S. Khalid, S. S. Almutairi, M. Abohashrh, S. A. Memon, J. Khan, Depression detection in social media: A comprehensive review of machine learning and deep learning techniques, IEEE Access 13 (2025) 12789–12818. URL: https://doi.org/10.1109/ACCESS.2025.3530862.
[6] J. Devlin, M. C. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423.
[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL: http://arxiv.org/abs/1907.11692.
[8] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778/.
[9] Z. Zhang, S. Chen, M. Wu, K. Q. Zhu, Psychiatric scale guided risky post screening for early detection of depression, in: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), 2022, pp. 5220–5226. URL: https://doi.org/10.24963/ijcai.2022/725.
[10] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory-II, Psychological Corporation, 1996.
[11] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual E5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672.

A. BDI-II Symptom Templates

The psychiatric-scale-guided screening stage compares each post against the 24 templates listed below. The first three are direct statements of depression, while the remaining 21 correspond to the individual symptom items of the BDI-II inventory.

1	I feel depressed.	13	I always cry.
2	I am diagnosed with depression.	14	I am hard to stay still.
3	I am treating my depression.	15	It's hard to get interested in things.
4	I feel sad.	16	I have trouble making decisions.
5	I am discouraged about my future.	17	I feel worthless.
6	I always fail.	18	I don't have energy to do things.
7	I don't get pleasure from things.	19	I have changes in my sleeping pattern.
8	I feel quite guilty.	20	I am always irritable.
9	I expected to be punished.	21	I have changes in my appetite.
10	I am disappointed in myself.	22	I feel hard to concentrate on things.
11	I always criticize myself for my faults.	23	I am too tired to do things.
12	I have thoughts of killing myself.	24	I have lost my interest in sex.