# LCDAA@eRisk 2026: Symptom-Level Evidence for Contextualized Early Detection of Depression

Notebook for the eRisk Lab at CLEF 2026

Cielo Aholiva Higuera-Gutiérrez[1,*,†], Sergio Mendez-Mendoza[1,†], Jesús Antonio Navarrete-López[1,†], Joan Manuel Raygoza-Romero[1,†], Juliette Ninnie Lazo-Vazquez[2,†], Irvin Hussein López-Nava[1,†] and Manuel Montes-y-Gómez[3,†]

[1]*Departamento de Ciencias de la Computación, Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE)*
[2]*Facultad de Ciencias, Universidad Autónoma de Baja California (UABC)*
[3]*Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)*

**Abstract**

This paper describes our participation in the CLEF eRisk 2026 Task 2, Contextualized Early Detection of Depression. We propose a symptom-oriented pipeline that models depression risk incrementally by combining three components: a template-based screening stage that estimates the semantic relevance of each writing with respect to BDI-II symptom descriptions, a dynamic buffer that retains the most symptom-relevant writings observed over time, and a set of 21 symptom-level classifiers that construct a structured user representation reflecting the clinical dimensions of depression. These features are used by a user-level classifier to issue early risk alerts. Additionally, a temporal Transformer-based approach is evaluated as a complementary strategy. The official results show that the symptom-oriented pipeline achieved the strongest decision-based F1 among our submissions and improved ranking performance as more evidence accumulated, while the Transformer-based runs produced earlier alerts. These findings suggest that explicitly modeling symptom-level evidence over time supports reliable and interpretable depression risk detection.

**Keywords**

Contextualized early detection, Depression detection, Social media, BDI-II symptoms, Symptom-oriented modeling, Risk prediction

## 1. Introduction

Depression is a common mental disorder whose impact extends beyond emotional distress, affecting daily functioning, interpersonal relationships, academic and work performance, and overall quality of life. Its clinical relevance is also reflected in its high prevalence, considering that depressive disorders affect approximately 5.7% of adults worldwide according to the World Health Organization (WHO) [1].

The onset of major depressive disorder frequently occurs between mid-adolescence and mid-adulthood, and nearly 40% of individuals experience their first episode before the age of 20 [2]. This early onset makes timely detection particularly critical, as delayed intervention is associated with worse long-term outcomes.

Beyond its prevalence, depression is characterized by a heterogeneous clinical presentation. Symptoms may span affective dimensions, such as persistent sadness or loss of interest; cognitive dimensions, such as feelings of worthlessness or difficulty concentrating; and somatic or behavioral changes, such as

*CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany*
*Corresponding author.
†These authors contributed equally.
aholiva@cicese.edu.mx (C. A. Higuera-Gutiérrez); smendez@cicese.edu.mx (S. Mendez-Mendoza); jnavarrete@cicese.edu.mx (J. A. Navarrete-López); jraygoza@cicese.edu.mx (J. M. Raygoza-Romero); juliette.lazo@uabc.edu.mx (J. N. Lazo-Vazquez); hussein@cicese.edu.mx (I. H. López-Nava); mmontesg@inaoep.mx (M. Montes-y-Gómez)
0009-0005-1172-2679 (C. A. Higuera-Gutiérrez); 0009-0009-2890-4822 (J. A. Navarrete-López); 0000-0003-3085-5678 (J. M. Raygoza-Romero); 0009-0002-6865-4468 (J. N. Lazo-Vazquez); 0000-0003-3979-9465 (I. H. López-Nava); 0000-0002-7601-501X (M. Montes-y-Gómez)

fatigue or changes in appetite [3, 4]. This variability makes early identification challenging, especially in non-clinical settings where symptoms are not systematically assessed. For this reason, standardized instruments such as the Beck Depression Inventory-II (BDI-II) organize depressive manifestations into symptom-based dimensions and provide a framework for assessing depressive severity [5].

In recent years, social media platforms have emerged as a source of user-generated data for mental health research. Users express experiences, struggles, and behavioral patterns through their posts, providing signals that may reflect aspects of their psychological state [6, 7]. Nevertheless, detecting depression from social media remains challenging because relevant signals are often sparse, noisy, and temporally distributed across multiple writings rather than concentrated in a single post.

The eRisk shared task series provides evaluation frameworks for the early detection of risk on the Internet. In this context, CLEF eRisk 2026 Task 2 addresses Contextualized Early Detection of Depression, where systems are required to process user activity sequentially and analyze full conversational threads rather than isolated posts [8, 9]. This setting is particularly relevant for social media depression detection, since risk-related evidence may be sparse, noisy, and distributed across multiple interactions.

In this work, we present our participation in this task and propose a symptom-oriented approach for contextualized early depression detection. Our system uses embedding-based similarity to identify symptom-relevant writings, retains the most informative evidence through a dynamic buffer, and constructs a structured user representation based on standardized depression symptom dimensions. These signals are then integrated into a user-level classifier to estimate depression risk and support early decision-making.

## 2. Related Work

Early risk detection has been a central objective of the eRisk shared tasks, which provide benchmark scenarios for identifying users at risk from online writings under incremental conditions. Unlike standard text classification settings, eRisk evaluates not only whether a system identifies users at risk, but also how early the decision is made. This has motivated approaches that model user activity over time and optimize both classification and early detection metrics [10, 11, 12].

Earlier eRisk editions addressed several mental health-related risks, including depression, pathological gambling, and anorexia, mainly from longitudinal histories authored by a single target user [13, 14, 12]. These tasks highlighted the importance of accumulating evidence across a user's writings, since risk-related signals are often sparse and may not appear in every post. In parallel, eRisk has also included symptom-oriented tasks, such as the "Search for Symptoms of Depression", where texts are ranked according to their relevance to the 21 symptoms defined by the Beck Depression Inventory (BDI) [5]. Although this formulation is not an early detection task, it provides a useful basis for representing depression through specific symptom dimensions rather than only as a global user-level label.

More recently, contextualized early detection has extended the traditional user-history setting by incorporating full conversational threads [12, 8, 9]. In this scenario, recent systems have explored different strategies to use or filter contextual information. Larrayoz et al. [15] proposed semantic relabeling and risk signal fusion between target-user and contextual scores. Zi et al. [16], building on the psychiatric scale-guided risky post screening approach proposed by Zhang et al. [17], introduced a learnable screening framework with a dynamic memory buffer to prioritize risk-indicative posts. Casamayor et al. [18] compared classical and long-context approaches, including SVM and Longformer models, showing that traditional classifiers can remain competitive when large textual histories must be processed efficiently.

Our work builds on these directions by adapting symptom-level evidence to the contextualized early detection setting. The proposed system identifies symptom-relevant writings, retains the most informative evidence through a dynamic buffer, and applies 21 BDI-II-inspired symptom-level classifiers to construct user-level features for early depression detection. We also compare this structured symptom-oriented representation with a temporal Transformer model over sequences of target-user embeddings.

## 3. Dataset and preprocessing

The dataset used for system development corresponds to the contextualized collection released for eRisk 2025 Task 2. It consists of conversational threads in JSON format, where each thread contains a submission and its associated comments, forming a hierarchical conversation through reply relations. Beyond textual content, the data include metadata such as author, timestamp, target subject, and round number, enabling the distinction between the target user's writings and the surrounding conversational context. For modeling purposes, the original nested JSON structure was reorganized into a tabular and chronological representation in which each row corresponds to a target subject at a given round. For each instance, we separated the text written by the target user from the contextual text written by other participants in the same thread, and used the round number and timestamp to preserve the temporal order of the data.

The main statistics of the collection are summarized in Table 1. The dataset contains 909 target subjects, 102 belonging to the depression group and 807 to the control group. Users in the depression group produce more threads per subject on average (397.7 vs. 295.9), span longer temporal periods (approximately 1,695 vs. 958 days), and generate comments that are more frequent and longer than those of control users.

**Table 1**
Main statistics of the Task 2 test collection.

| Statistic | Depression | Control |
|---|---|---|
| Number of subjects | 102 | 807 |
| Number of threads | 40,563 | 238,033 |
| Average number of threads per subject | 397.7 | 295.9 |
| Average number of days from first to last thread | $\approx$ 1695 | $\approx$ 958 |
| Average number of comments per thread | 65.1 | 44.6 |
| Average number of words per comment | 33.8 | 25.6 |

## 4. Methodology

### 4.1. Temporal Transformer-based pipeline

The first strategy explored in this work is a temporal Transformer-based approach, illustrated in Figure 1. Specifically this strategy was designed to model the evolution of the target user's writings over time. Instead, each target user is represented as a chronological sequence of round-level text embeddings.

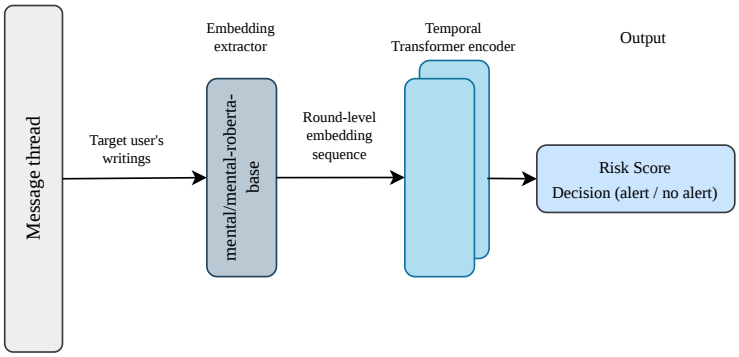


**Figure 1:** Overview of the temporal Transformer-based approach.

#### 4.1.1. Round-level embedding representation

For each round, the text written by the target user was encoded using `mental/mental-roberta-base`. The embedding for each writing was obtained by applying mean pooling over the last hidden states using the attention mask, followed by normalization. This produced one dense vector representation per target-user writing.

For a target user $u$, the chronological sequence of embeddings is represented as:

$$X_u = [\mathbf{e}_{u,1}, \mathbf{e}_{u,2}, \ldots, \mathbf{e}_{u,T_u}] \tag{1}$$

where $\mathbf{e}_{u,t}$ is the embedding of the target user's writing at round $t$, and $T_u$ is the number of available writings for that user. Since users have different numbers of writings, sequences were padded or truncated to a fixed maximum length $T_{\max}$. A padding mask was used to ensure that the model attended only to real writings and ignored padded positions.

#### 4.1.2. Transformer sequence model

The temporal classifier receives the padded sequence of round-level embeddings and a corresponding padding mask. The 768-dimensional embeddings are first projected to a lower-dimensional space via a dense layer with ReLU activation and dropout, then processed by Transformer encoder blocks with multi-head self-attention, residual connections, and feed-forward layers. The padding mask prevents padded positions from contributing to the attention computation. Masked mean pooling is applied across the temporal dimension to obtain a single user-level vector, which is passed through a dense layer and a sigmoid output unit to produce the depression risk score.

The model was trained with binary cross-entropy loss, class weights to address class imbalance, and early stopping based on validation loss.

#### 4.1.3. Early inference strategy

During inference, the Transformer model was applied incrementally to simulate the early detection scenario. At each round $t$, the model receives only the embeddings observed up to that point:

$$X_{u,t} = [\mathbf{e}_{u,1}, \mathbf{e}_{u,2}, \ldots, \mathbf{e}_{u,t}] \tag{2}$$

The sequence is padded to $T_{\max}$, and the corresponding mask is updated. The model then produces a risk score $p_{u,t}$, which is used as the submitted depression score. A binary decision is obtained by comparing this score with a threshold $\tau$. The threshold was selected on the validation set according to the early detection criterion used for the run.

Two Transformer configurations were used in the final submissions, differing in the maximum number of rounds considered by the model. One configuration used $T_{\max} = 300$, while the other used $T_{\max} = 400$. This allowed us to analyze whether providing a longer user history improved the balance between classification performance and early detection latency.

### 4.2. Symptom-oriented pipeline

The second strategy explored in this work is a symptom-oriented early detection pipeline, illustrated in Figure 2. Unlike the temporal Transformer approach, which models the target user's writings directly as a sequence of embeddings, this pipeline explicitly identifies and accumulates symptom-relevant evidence over time. At each round, the system processes the available information for a target subject and updates a user-level representation based on writings that are semantically related to depressive manifestations.

The pipeline consists of four main stages: symptom-template similarity screening, dynamic buffering of relevant writings, symptom-level representation, and user-level classification for early decision-making.

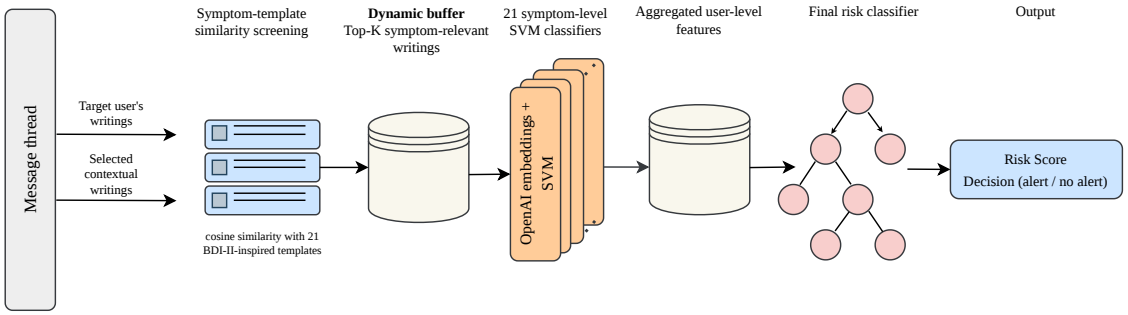


**Figure 2:** Overview of the symptom-oriented pipeline proposed for contextualized early depression detection.

### 4.2.1. Symptom-template similarity screening

The first stage of the system estimates the semantic proximity between each screening input and a set of symptom-oriented reference statements. This stage was motivated by previous risk-post screening strategies, where descriptions derived from psychological scales are used to prioritize writings that are more likely to contain risk-relevant evidence [16]. In our implementation, the reference set consists of 21 templates derived from the BDI-II symptom dimensions [5]. The complete set of templates is provided in Appendix A.

At each round, the system constructs a screening input by combining the target user's writing with selected contextual information from the same thread. The contextual component is included only when it is likely to be related to the target user, using reply-chain information and simple linguistic cues such as second-person references and support-oriented expressions. This allows the screening stage to consider the user's own writing together with relevant surrounding interaction, while reducing unrelated conversational noise.

Let $x_i$ denote the screening input at round $i$, and let $Q = \{q_1, \ldots, q_{21}\}$ be the set of symptom-oriented templates. Both the screening input and the templates were encoded using `text-embedding-3-small`. The semantic relevance of $x_i$ with respect to each template was computed using cosine similarity:

$$s_{ij} = \cos(e(x_i), e(q_j)), \quad j = 1, \ldots, 21. \tag{3}$$

This produces a 21-dimensional template-similarity vector $\mathbf{s}_i = [s_{i1}, s_{i2}, \ldots, s_{i21}]$ for each screening input. From this vector, we derived three global screening summaries: the maximum similarity score, the mean similarity across all templates, and the mean of the three highest similarity scores. These values are used to estimate the symptom-relevance of the current writing and to prioritize writings for the dynamic buffer.

### 4.2.2. Dynamic buffer

Depression-related evidence in social media is often sparse and unevenly distributed across a user's writings. To reduce the influence of unrelated or weakly informative content, the system maintains a dynamic buffer for each target user. This buffer retains the target user's writings that have shown the highest symptom relevance up to the current round.

The buffer is updated incrementally after each round. For each new target-user writing, a screening score is obtained from the symptom-template similarity stage. This score corresponds to the maximum similarity between the screening input and the symptom templates. The screening input may include both the target user's writing and selected contextual information from the same thread, but the buffer stores the target user's writing as the main unit of evidence.

For a user $u$ at round $t$, the buffer $B_{u,t}$ stores the top $K$ target-user writings with the highest screening

scores among those observed up to that point:

$$B_{u,t} = \text{TopK}\left(\{(x_i, r_i) : i \leq t\}, K\right) \tag{4}$$

where $x_i$ denotes a target-user writing and $r_i$ is its corresponding screening score. In our implementation, the buffer size was set to $K = 5$. When the buffer has available capacity, the new writing is added directly. Once the buffer is full, the new writing replaces the least relevant buffered entry only if its screening score is higher.

This mechanism allows the system to accumulate a compact set of symptom-relevant evidence over time. The retained writings are then used in the next stage to construct a symptom-level representation of the target user.

#### 4.2.3. Symptom-level representation

After the dynamic buffer is updated, the retained writings are transformed into a symptom-level representation. This stage aims to characterize the type of depression-related evidence in the writings selected by the buffer, rather than relying only on their semantic proximity to symptom templates.

For this purpose, we used 21 independent symptom-level classifiers, each associated with one symptom dimension derived from the Beck Depression Inventory-II (BDI-II). These classifiers were trained offline using the "Search for Symptoms of Depression" datasets released in previous eRisk evaluation campaigns [13]. These datasets provide sentence-level annotations indicating whether a sentence contains evidence related to each depressive symptom. Therefore, each symptom was modeled as an independent binary classification problem.

Each symptom classifier was implemented as a Support Vector Machine (SVM) trained on dense text embeddings obtained from an OpenAI embedding model. In this setting, the embedding model maps each input text into a semantic vector representation, while the SVM learns to identify evidence associated with a specific symptom. The resulting models were used as feature extractors in the contextualized early detection task.

Given a target-user writing $x_i$ retained in the buffer, its embedding representation $\mathbf{e}_i$ is first computed. Then, each symptom classifier $f_j$, with $j = 1, \ldots, 21$, produces a score indicating the presence of evidence for symptom $j$:

$$z_{ij} = f_j(\mathbf{e}_i) \tag{5}$$

The outputs of the 21 classifiers are concatenated to obtain a symptom vector for the writing:

$$\mathbf{z}_i = [z_{i1}, z_{i2}, \ldots, z_{i21}]. \tag{6}$$

In our pipeline, this process is applied only to the target-user writings stored in the dynamic buffer. This reduces the amount of text processed by the symptom-level models and focuses the representation on the writings previously identified as more symptom-relevant by the screening stage. The surrounding conversational context is not directly classified by the 21 symptom-level models; instead, it is incorporated through the screening input and through contextual indicators used in the final feature vector.

To construct a user-level representation at each round, the symptom vectors of the buffered writings are aggregated. For each symptom, we compute summary statistics such as the mean and maximum score across the buffer. The mean summarizes the overall presence of a symptom across the selected writings, whereas the maximum captures strong isolated evidence that may appear in a single writing. The resulting aggregated representation summarizes the symptom-level evidence accumulated for the target user up to the current round.

#### 4.2.4. User-level classification and early decision

At each round, a user-level feature vector is constructed by combining screening information, buffer evidence, symptom-level scores, and contextual indicators. Let $\mathbf{s}_{u,t}$ denote the symptom-template

similarity vector computed for user $u$ at round $t$, and let $\mathbf{g}_{u,t}$ denote its screening summaries, including the maximum similarity, the mean similarity, and the mean of the three highest similarity scores.

The dynamic buffer provides accumulated evidence from the writings retained up to round $t$. From this buffer, we extract aggregated screening features, denoted as $\mathbf{b}_{u,t}^{scr}$, and aggregated symptom-level scores produced by the 21 symptom classifiers, denoted as $\mathbf{b}_{u,t}^{sym}$. Contextual information is represented by a vector $\mathbf{c}_{u,t}$, which includes indicators such as the presence of contextual text, second-person references, and support-oriented expressions.

The final representation of user $u$ at round $t$is defined as:

$$\mathbf{h}_{u,t} = [\mathbf{s}_{u,t}; \mathbf{g}_{u,t}; \mathbf{b}_{u,t}^{scr}; \mathbf{b}_{u,t}^{sym}; \mathbf{c}_{u,t}] \tag{7}$$

where $\mathbf{h}_{u,t}$ summarizes the current symptom-template similarity, the accumulated evidence stored in the dynamic buffer, the symptom-level evidence, and the contextual information available at that round.

These features are used as input to a user-level classifier that estimates the depression risk of the target subject at each round. Several machine learning classifiers were evaluated, including Random Forest, Extra Trees, Histogram Gradient Boosting, Gradient Boosting, Passive-Aggressive, and CatBoost.

For each user $u$ and round $t$, the classifier $F$ produces a continuous risk score:

$$p_{u,t} = F(\mathbf{h}_{u,t}) \tag{8}$$

A binary decision is then obtained by comparing this score with a decision threshold $\tau$:

$$\hat{y}_{u,t} = \begin{cases} 1, & \text{if } p_{u,t} \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Following the official early detection protocol, $\hat{y}_{u,t} = 1$ represents an alert and is considered final for that user, whereas $\hat{y}_{u,t} = 0$ indicates no alert and allows the system to issue an alert in a later round. After an alert is emitted, the system continues processing subsequent writings and submitting updated risk scores, since these scores are also used for ranking-based evaluation.

## 5. Results and discussion

Table 2 reports the internal evaluation results for the user-level classifiers in the symptom-oriented pipeline. Among the models evaluated with the default threshold, Random Forest achieved the highest macro-F1, while CatBoost obtained the lowest ERDE5, revealing a trade-off between classification quality and detection timeliness. Model selection was performed on the validation set using a combined criterion that prioritized ERDE5, followed by Macro-F1, latencyTP, and recall macro. Under this criterion, Histogram Gradient Boosting (HGB) was selected as the main classifier. A threshold search was then conducted for HGB to define the tuned-threshold configuration ($\tau = 0.35$), which reduced ERDE5 and improved F-latency compared with the default threshold. These two configurations correspond to Run 0 (HGB, tuned threshold $\tau = 0.35$) and Run 1 (HGB, default threshold $\tau = 0.50$) in the official submission.

Table 3 reports the internal results for the temporal Transformer models. Both configurations achieved strong classification performance and notably low detection latency. The $T_{\max} = 400$ model obtained the highest macro-F1 (0.900), while the $T_{\max} = 300$ model yielded lower ERDE5 and higher F-latency. These configurations correspond to Run 3 ($T_{\max} = 300$) and Run 4 ($T_{\max} = 400$) in the official submission.

## 6. Run configurations

Based on the development-stage analysis, five runs were submitted to the official eRisk 2026 Task 2 server. The runs were designed to compare two different strategies for contextualized early depression

**Table 2**
Internal test results of the symptom-oriented user-level classifiers.

| Model | P | R | F1-Macro | ERDE5 | latencyTP | F-latency |
|---|---|---|---|---|---|---|
| RF | 0.783 | 0.900 | 0.860 | 0.096 | 77.389 | 0.635 |
| CatBoost | 0.513 | 1.000 | 0.854 | 0.068 | 59.950 | 0.568 |
| ExtraTrees | 0.739 | 0.850 | 0.840 | 0.099 | 83.118 | 0.612 |
| GB | 0.581 | 0.900 | 0.834 | 0.081 | 71.000 | 0.600 |
| HGB | 0.600 | 0.900 | 0.833 | 0.086 | 79.944 | 0.559 |
| HGB tuned | 0.682 | 0.938 | 0.789 | 0.063 | 66.267 | 0.687 |
| Passive-Aggressive | 0.172 | 1.000 | 0.766 | 0.121 | 22.350 | 0.258 |

**Table 3**
Internal test results of the temporal Transformer models.

| Model | P | R | F1-Macro | ERDE5 | latencyTP | F-latency |
|---|---|---|---|---|---|---|
| Transformer $T_{\max} = 300$ | 0.833 | 0.750 | 0.883 | 0.050 | 2.000 | 0.806 |
| Transformer $T_{\max} = 400$ | 0.842 | 0.800 | 0.900 | 0.050 | 2.000 | 0.770 |

detection. The first three runs were based on the symptom-oriented pipeline, while the last two runs used the temporal Transformer-based approach.

For the symptom-oriented strategy, we submitted two configurations based on the Histogram Gradient Boosting classifier and one configuration based on Random Forest. The first run used the HGB classifier with a threshold selected on the validation set, whereas the second run used the same classifier with the default threshold. The third run used Random Forest with the default threshold, as this model showed strong decision-based performance during development.

For the Transformer-based strategy, two runs were submitted using different maximum sequence lengths. These runs were included to evaluate whether extending the temporal window of user writings improves the balance between classification performance and early detection behavior.

Table 4 summarizes the submitted configurations.

**Table 4**
Submitted run configurations for eRisk 2026 Task 2.

| Run | Approach | Model | Configuration |
|---|---|---|---|
| 0 | Symptom-oriented pipeline | HGB | Tuned threshold, $\tau = 0.35$ |
| 1 | Symptom-oriented pipeline | HGB | Default threshold, $\tau = 0.50$ |
| 2 | Symptom-oriented pipeline | RF | Default threshold, $\tau = 0.50$ |
| 3 | Temporal Transformer | Transformer | $T_{\max} = 300$ |
| 4 | Temporal Transformer | Transformer | $T_{\max} = 400$ |

## 6.1. Official results

Table 5 presents the official decision-based results, including selected top-performing systems and the five LCDAA runs. Among our submissions, Run 2 achieved the best decision-based performance, with an F1-score of 0.78 and balanced precision and recall. This result was 0.05 points below the best F1-score reported in the official table, obtained by HUGETIME, and remained close to other competitive systems. Run 2 corresponds to the symptom-oriented pipeline using Random Forest, suggesting that the proposed symptom-level representation provided competitive user-level classification performance.

Runs 0 and 1, both based on the HGB classifier, obtained similar performance. Both configurations achieved the same recall, while Run 1 obtained slightly higher precision, F1-score, and latency-weighted F1 than the tuned-threshold configuration in Run 0.. These results suggest that threshold adjustment

affected the precision–recall balance but did not improve the official performance compared with the default threshold.

The Transformer-based runs showed a different behavior. Run 3 obtained the lowest ERDE5 among our submissions and the earliest detection latency (latencyTP $= 1.00$, speed $= 1.00$), matching the strongest latency values reported by several systems, but at the cost of a lower F1-score. Run 4 improved F1 with respect to Run 3 while preserving low latency. Overall, the decision-based results indicate a trade-off between classification strength and early alerting across the submitted strategies.

**Table 5**
Decision-based evaluation for Task 2. The table includes selected top-performing teams and all LCDAA runs.

| Team | Run | P | R | F1 | ERDE5 | ERDE50 | latencyTP | Speed | F-latency |
|---|---|---|---|---|---|---|---|---|---|
| HUGETIME | 0 | 0.80 | 0.87 | 0.83 | 0.15 | 0.05 | 9.00 | 0.97 | 0.81 |
| UNED-GELP | 2 | 0.79 | 0.85 | 0.82 | 0.11 | 0.06 | 4.00 | 0.99 | 0.81 |
| NoirByte | 0 | 0.88 | 0.76 | 0.82 | 0.15 | 0.09 | 20.00 | 0.93 | 0.76 |
| MindAILab | 0 | 0.84 | 0.80 | 0.82 | 0.16 | 0.08 | 16.00 | 0.94 | 0.77 |
| INSA-Lyon | 0 | 0.76 | 0.85 | 0.80 | 0.16 | 0.07 | 13.00 | 0.95 | 0.76 |
| LCDAA | 0 | 0.66 | 0.81 | 0.73 | 0.11 | 0.07 | 3.00 | 0.99 | 0.72 |
| LCDAA | 1 | 0.69 | 0.81 | 0.74 | 0.11 | 0.07 | 4.00 | 0.99 | 0.74 |
| LCDAA | 2 | 0.78 | 0.78 | 0.78 | 0.13 | 0.08 | 10.00 | 0.96 | 0.75 |
| LCDAA | 3 | 0.70 | 0.70 | 0.70 | 0.09 | 0.07 | **1.00** | **1.00** | 0.70 |
| LCDAA | 4 | 0.72 | 0.74 | 0.73 | 0.11 | 0.07 | 2.00 | **1.00** | 0.73 |

Table 6 summarizes the official ranking-based results and places the LCDAA runs in context with selected top-performing systems. The Transformer-based runs obtained the strongest ranking performance among our submissions, with Run 4 reaching an NDCG@100 of 0.73 after one writing, close to the best value reported in the table. This suggests that temporal embedding representations can provide useful risk estimates even with limited evidence.

As more writings became available, the symptom-oriented runs achieved better ranking performance. Runs 0, 1, and 2 reached an NDCG@100 of 0.84 after 100 writings and 0.88 after 500 writings, approaching the strongest systems, whose NDCG@100 at 500 writings ranged from 0.89 to 0.91. These results show complementary behavior between the two strategies: Transformer-based runs were more effective at the earliest stage, while symptom-oriented runs became more reliable as more evidence was observed.

**Table 6**
Ranking-based evaluation for Task 2. The table includes selected top-performing teams and all LCDAA runs.

| Team | Run | 1 writing | | | 100 writings | | | 250 writings | | | 500 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 | P@10 | NDCG@10 | NDCG@100 |
| DeepCare | 0 | **1.00** | **1.00** | **0.74** | **1.00** | **1.00** | **0.88** | **1.00** | **1.00** | **0.90** | **1.00** | **1.00** | **0.91** |
| MindAILab | 0 | **1.00** | **1.00** | 0.69 | **1.00** | **1.00** | 0.86 | **1.00** | **1.00** | 0.89 | **1.00** | **1.00** | 0.90 |
| NoirByte | 0 | **1.00** | **1.00** | 0.72 | **1.00** | **1.00** | 0.86 | **1.00** | **1.00** | 0.88 | **1.00** | **1.00** | 0.89 |
| INSA-Lyon | 0 | 0.80 | 0.87 | 0.64 | **1.00** | **1.00** | 0.84 | **1.00** | **1.00** | 0.87 | **1.00** | **1.00** | 0.89 |
| HUTECH-NLP | 0 | **1.00** | **1.00** | 0.71 | **1.00** | **1.00** | 0.82 | **1.00** | **1.00** | 0.84 | **1.00** | **1.00** | 0.85 |
| LCDAA | 0 | 0.90 | 0.92 | 0.61 | **1.00** | **1.00** | 0.84 | **1.00** | **1.00** | 0.87 | 0.90 | 0.94 | 0.88 |
| LCDAA | 1 | 0.90 | 0.92 | 0.61 | **1.00** | **1.00** | 0.84 | **1.00** | **1.00** | 0.87 | 0.90 | 0.94 | 0.88 |
| LCDAA | 2 | 0.90 | 0.94 | 0.63 | **1.00** | **1.00** | 0.84 | **1.00** | **1.00** | 0.86 | 0.90 | 0.94 | 0.88 |
| LCDAA | 3 | 0.90 | 0.92 | 0.71 | 0.90 | 0.93 | 0.79 | 0.90 | 0.94 | 0.79 | 0.90 | 0.88 | 0.84 |
| LCDAA | 4 | 0.90 | 0.93 | 0.73 | 0.90 | 0.93 | 0.78 | 0.90 | 0.92 | 0.79 | 0.90 | 0.86 | 0.81 |

Overall, the official results show complementary strengths between the two families of approaches.

The symptom-oriented pipeline, particularly the Random Forest configuration, achieved the strongest decision-based F1 and the best ranking performance after evidence accumulation, demonstrating that explicitly modeling BDI-II symptom dimensions over time provides a competitive and interpretable alternative to end-to-end neural approaches. In contrast, the Transformer-based runs produced earlier alerts and stronger ranking performance at the first-writing checkpoint, but their advantage decreased as more writings became available.

### 6.2. Discussion

The symptom-oriented pipeline achieved the strongest decision-based performance among our submissions, particularly when using Random Forest. This suggests that combining template-based screening, dynamic buffering, and symptom-level features provides useful evidence for user-level depression classification. This behavior may be explained by the accumulation of evidence across symptom dimensions: instead of relying on a single depression score, the system summarizes the presence of multiple depressive manifestations over time, constructing a structured user profile that reflects the clinical heterogeneity of depression. Beyond predictive performance, this representation offers an additional advantage: each dimension of the feature vector corresponds to a specific BDI-II symptom, making it possible to inspect which symptoms contributed most to a given decision. This degree of interpretability is not directly available in the Transformer-based approach, where the depression risk score is produced by an opaque sequence model.

In contrast, the Transformer-based runs were more effective for early alerting. They produced the lowest latency values among our submissions, but this came at the cost of lower decision-based F1. This indicates that the temporal models were more sensitive to early signals, although less stable in terms of final classification performance.

The ranking-based evaluation also suggests that both approaches capture different aspects of the task. Transformer-based runs performed better at the earliest checkpoint, whereas the symptom-oriented runs improved as more writings became available. Overall, these findings point to a trade-off between early sensitivity and accumulated symptom-based evidence. Future work should explore hybrid strategies that combine temporal modeling with explicit symptom-level representations.

## 7. Conclusion and future work

This paper described our participation in CLEF eRisk 2026 Task 2, Contextualized Early Detection of Depression. We explored two strategies for early depression detection: a symptom-oriented pipeline and a temporal Transformer-based approach. The symptom-oriented pipeline combines template-based screening, dynamic buffering, and 21 BDI-II-inspired symptom-level classifiers to construct a structured user representation that is updated as new writings become available. This representation explicitly models different depressive manifestations over time, making the decision process more interpretable than approaches based only on a single global risk score.

The official results showed complementary behavior between the two strategies. The symptom-oriented pipeline achieved the strongest decision-based performance among our submissions, particularly when using Random Forest, and improved its ranking performance as more evidence accumulated. In contrast, the Transformer-based runs were more effective for early alerting, obtaining lower latency values. These findings suggest that symptom-level evidence can support robust user-level depression classification, while temporal embedding models may capture early risk signals more rapidly.

Two main directions are identified for future work. First, the symptom-level classifiers are currently applied only to the writings retained in the dynamic buffer due to computational constraints. Applying these models to a broader set of writings, and potentially to selected contextual information, could provide a richer symptom-level representation and capture evidence that the current screening stage may discard. Second, the results showed that the decision threshold selected during development did not generalize optimally to the official sequential evaluation. Exploring probability calibration and

adaptive thresholding techniques could improve threshold stability and produce more reliable risk scores across different evaluation conditions.

## 8. Acknowledgments

This research was financially supported by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through the scholarship grant 4034491.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for grammar and spelling checks. After using this tool, the authors carefully 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: January 21, 2026.
[2] G. S. Malhi, J. J. Mann, Depression: Seminar, The Lancet 392 (2018) 10161.
[3] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed., American Psychiatric Publishing, Washington, DC, 2013. doi:10.1176/appi.books.9780890425596.
[4] D. Goldberg, The heterogeneity of "major depression", World Psychiatry 10 (2011) 226.
[5] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory-II, Psychological Corporation, San Antonio, TX, 1996.
[6] P. Dodemaide, M. Merolli, N. Hill, L. Joubert, Do social media impact young adult mental health and well-being? a qualitative study, The British Journal of Social Work 52 (2022) 4664–4683.
[7] C. R. Blease, Too many 'friends,'too few 'likes'? evolutionary psychology and 'facebook depression', Review of General Psychology 19 (2015) 1–13.
[8] 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.
[9] 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.
[10] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: experimental foundations, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 346–360.
[11] F. Crestani, D. E. Losada, J. Parapar, Early detection of mental health disorders by social media monitoring, Studies in Computational Intelligence 1018 (2022) 4.
[12] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Early risk prediction on the internet, in: International conference of the cross-language evaluation forum for European languages, Springer, 2025, pp. 242–265.
[13] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: International conference of the cross-language evaluation forum for European languages, Springer, 2023, pp. 294–315.

[14] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II, Springer, 2024, pp. 73–92.
[15] X. Larrayoz, A. Casillas, A. Pérez, Leveraging conversational context and semantic relabeling for early depression detection, Working Notes of CLEF (2025) 9–12.
[16] Y. Zi, B. Wang, Y. Zhao, B. Qin, 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, Working Notes of CLEF (2025) 9–12.
[17] Z. Zhang, S. Chen, M. Wu, K. Q. Zhu, Psychiatric scale guided risky post screening for early detection of depression, arXiv preprint arXiv:2205.09497 (2022).
[18] A. Casamayor, V. Ahuir, A. Molina, L.-F. Hurtado, Elirf-upv at erisk 2025: New approaches to the detection and early detection of symptoms and signs of depression, Working Notes of CLEF (2025).

# A. Appendices

## A.1. Symptom templates

Table 7 presents the 21 symptom-oriented templates used in the symptom-template similarity screening stage. Each template was encoded using the same embedding model as the user writings and compared through cosine similarity.

Table 7: Symptom-oriented templates used for embedding-based screening.

| No. | Template |
|---|---|
| 1 | I feel sad. |
| 2 | I am discouraged about my future. |
| 3 | I always fail. |
| 4 | I do not get pleasure from things. |
| 5 | I feel guilty. |
| 6 | I expected to be punished. |
| 7 | I am disappointed in myself. |
| 8 | I criticize myself for my faults. |
| 9 | I have thoughts of killing myself. |
| 10 | I always cry. |
| 11 | I am too restless to stay still. |
| 12 | It is hard to get interested in things. |
| 13 | I have trouble making decisions. |
| 14 | I feel worthless. |
| 15 | I do not have energy to do things. |
| 16 | I have changes in my sleeping pattern. |
| 17 | I am always irritable. |
| 18 | I have changes in my appetite. |
| 19 | I find it hard to concentrate on things. |
| 20 | I am too tired to do things. |
| 21 | I have lost interest in sex. |