UNED-GELP at eRisk 2026: Temporal Architectures for Depression Detection

eRisk at CLEF 2026

Jorge Fernandez-Hernandez[1], Juan Martinez-Romo[1,2,†] and Lourdes Araujo[1,2,†]

[1]*NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain*

[2]*IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain*

Abstract

This paper describes our participation in the eRisk 2026 shared task at CLEF, specifically Task 2, focusing on the early detection of depression in social media environments. We propose a novel Transformer-based architecture, the Temporal Hierarchical Transformer, designed to combine textual representations with explicit temporal information while effectively overcoming the standard 512-token sequence length constraint inherent to traditional models. This approach proves highly robust in handling the dense publication frequency and continuous accumulation of posts typical of social media users. Our framework demonstrated competitive performance during the official evaluation phase, ultimately securing the second position in the laboratory's global leaderboard, thereby validating the critical role of time-aware hierarchical modeling for automated mental health screening.

Keywords

Depression detection, Temporal information, Early detection

1. Introduction

Mental health disorders currently represent one of the most widespread public health challenges worldwide. According to estimates by the World Health Organization (WHO) [1], approximately 970 million people suffer from some form of mental disorder, with depression and anxiety being the most prevalent conditions. Consequently, advancing research into early detection mechanisms is of paramount importance. In this regard, social media platforms play a pivotal role, as they provide a vast repository of longitudinal textual data. These digital traces enable the continuous monitoring of user-generated content over time, thereby facilitating the early identification of risk factors and symptomatic behavioral patterns.

In this context, initiatives such as eRisk play a crucial part by fostering the development of computational systems specialized in the early detection of various mental health disorders. These shared tasks not only drive technological innovation within the digital health domain but also provide a significant contribution by releasing new longitudinal datasets. These resources are indispensable for advancing scientific research and refining current predictive models.

This paper describes our participation in the eRisk 2026 shared task, specifically Task 2, which focuses on the early detection of depression on social media by contextualizing user messages [2, 3]. To solve this task, we explored several approaches specifically focused on leveraging the temporal information of messages. Our work is driven by the hypothesis that user temporal patterns are closely correlated with depressive symptoms, thus representing a key feature for the automated detection of the disorder.

The rest of the paper is structured as follows: in Section 2 we describe other related tasks and existing systems developed for early detection of mental disorders through text. In Section 3 we describe the

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

*Corresponding author.

†These authors contributed equally.

jfernandez@lsi.uned.es (J. Fernandez-Hernandez); juaner@lsi.uned.es (J. Martinez-Romo); lurdes@lsi.uned.es (L. Araujo)

0009-0001-8996-8303 (J. Fernandez-Hernandez); 0000-0002-6905-7051 (J. Martinez-Romo); 0000-0002-7657-4794 (L. Araujo)

task and our methods developed to solve it. In Section 4 we analyze and discuss the results obtained by our systems.

2. Related Work

Since its inaugural year, the eRisk lab has been at the forefront of exploring the early detection of various mental disorders through the analysis of digital traces. Throughout its successive editions, the initiative has addressed pathologies with distinct linguistic manifestations, including depression [4, 5, 6, 7, 8, 9, 10, 11], eating disorders [5, 8, 9, 10], or pathological gambling [7, 8, 9]. These tasks have consolidated a longitudinal evaluation methodology that currently serves as a benchmark in the field of computational mental health.

Since its introduction, the Transformer architecture has established itself as the state of the art in natural language processing (NLP). This trend directly extends to eRisk challenges, where classification methodologies have undergone a progressive evolution. Initially, systems relied on general-purpose models such as BERT [12]; however, subsequent research has shifted toward domain-specific Transformers tailored for the psychological and mental health fields, such as MentalBERT or MentalROBERTA [13]. Furthermore, approaches focused on deriving dense vector representations via Sentence Transformers have proven highly effective in tasks that move away from strict absolute classification, such as symptom extraction based on the BDI questionnaire or severity measurement in eating disorders [14, 15]. Lastly, hybrid methodologies have also been explored, leveraging pre-trained models for text feature extraction alongside LSTM networks to capture the sequential and temporal dynamics of social media posts [16]. In parallel, modeling complex social interactions and synthesizing summaries across multiple historical posts have been successfully integrated into classic classification challenges. These contextual frameworks are leveraged to process continuous dialogue histories, ultimately enhancing performance during the fine-tuning of discriminative Transformer architectures [17].

Medical literature has long established an intrinsic link between depression and the disruption of circadian rhythms [18], characterized by a significant tendency toward insomnia and delayed sleep phases in affected individuals. Studies leveraging large-scale social media data have identified distinct nocturnal activity patterns among individuals with depression, characterized by significantly heightened posting frequency during evening hours (19:00–00:00) and increased ruminative or negative content during pre-dawn periods [19, 20]. These sleep-wake and circadian features have shown high predictive utility, with daily phase delays serving as strong indicators for the onset of depressive episodes [21, 22]

3. Task 2: Contextualized Early Detection of Depression

This task is centered around the early detection of depressive symptoms on social media. A key distinction of this task compared to others in the field is the provision of the full situational context in which each message was authored. Specifically, the dataset consists of complete Reddit posts, including both the user-generated text and the context in which the message was published. The aim of this approach is to facilitate a more comprehensive analysis by capturing not only the individual's output but also the conversational threads surrounding each post.

After evaluating several architectures designed to integrate user content with situational context, we ultimately decided to focus the analysis exclusively on the user's messages, disregarding the external context. To construct a structured chronological timeline of each individual's activity, we systematically processed every post in which the user participated. For each indexed post, a unified textual representation was generated by concatenating all submissions authored by the user within that post. During this process, messages containing fewer than two words were excluded, as well as deleted Reddit messages that appeared as [Removed] after scraping. In cases where the post was originally authored by the user under analysis, both the title and the post body were explicitly concatenated into the final string. Consequently, our dataset is formally defined as $\mathcal{D} = \{U_1 : [(t_{1-1}, d_{1-1}), \ldots, (t_{1-n}, d_{1-n})], \ldots, U_{909} : [(t_{909-1}, d_{909-1}), \ldots, (t_{909-m}, d_{909-m})]\}$

Run	Architecture
0	Temporal Hierarchical Transformer with hour and minute
1	Temporal Hierarchical Transformer with time elapsed
2	Ensemble majority voting

Table 1
Summary of the runs with the architecture used.

where U_x denotes the x-th user who has engaged across a series of distinct n posts. Each post consists of a two-element tuple: a text string (t_{x-i}) formed by concatenating all of the user's messages within that specific post, alongside a timestamp (d_{x-i}) representing the exact date and time of their final message in that post.

To address Task 2, we proposed five distinct methodologies. However, two of the proposed systems did not achieve satisfactory results during the evaluation phase due to a software bug in the utilized library; therefore, they will not be discussed in detail. The system configurations for the remaining three runs are detailed in Table 1. The first two methodologies introduce a custom architecture termed Temporal Hierarchical Transformers. This approach builds upon the longer input sizes of the hierarchical baseline and novelly integrates temporal information. This integration is grounded in clinical evidence linking mood disorders to disruptions in circadian rhythms—a pathology that manifests digitally through a measurable increase in posting frequency during late-night hours. Finally, the third approach consists of an ensemble model based on a majority voting scheme, integrating the three individual methodologies that exhibited the most optimal performance during our preliminary experimental phases: both developed variants of the Temporal Hierarchical Transformer and one of the systems with the bug in evaluation phase.

Lastly, to adapt to the sequential, round-based evaluation pipeline that characterizes the testing phase in the eRisk lab, we implemented a dynamic modeling strategy based on sliding windows. Under this framework, instead of retaining an individual's entire and unbounded historical log, each user's profile is constrained to a fixed window length N. While the conventional update policy for such mechanisms typically limits retention to the most recent N chronological messages, during our preliminary experimental phases we explored various window sizes along with alternative update strategies. Among these, we evaluated a selective retention approach that preserved the N posts exhibiting the highest levels of negative sentiment or sadness indicators. Ultimately, after jointly balancing predictive performance, execution latency, and computational costs, we opted to configure the final system using a sliding window strictly comprising the user's last 50 active rounds.

3.1. Temporal Hierarchical Transformers: Run 0

Hierarchical Transformers are an evolution of the standard Transformer-based architecture, specifically designed to overcome the 512-token input limitation [23]. The core principle of these models involves partitioning the input text into shorter fragments prior to processing. Each segment is encoded into vector representations (embeddings), which are subsequently combined by integrating positional embedding information. This structure enables the capture of long-range contextual relationships more efficiently than other alternatives, such as Longformers, achieving superior results in long-document classification tasks.

The Temporal Hierarchical Transformer builds upon the foundational architecture of the standard hierarchical model but introduces a granularity centered on user chronology. In this approach, instead of segment texts sequentially, the model processes each individual post as a distinct unit, leveraging its associated timestamp. This design is informed by the clinical observation that sleep disturbances are often symptomatic of depression. Consequently, the system is designed to detect distinctive temporal behavioral patterns, such as a higher propensity for late-night or early-morning activity, which may serve as a discriminative feature between positive and negative cases.

To integrate the temporal dimension, the timestamps (hour and minute of publication) are processed

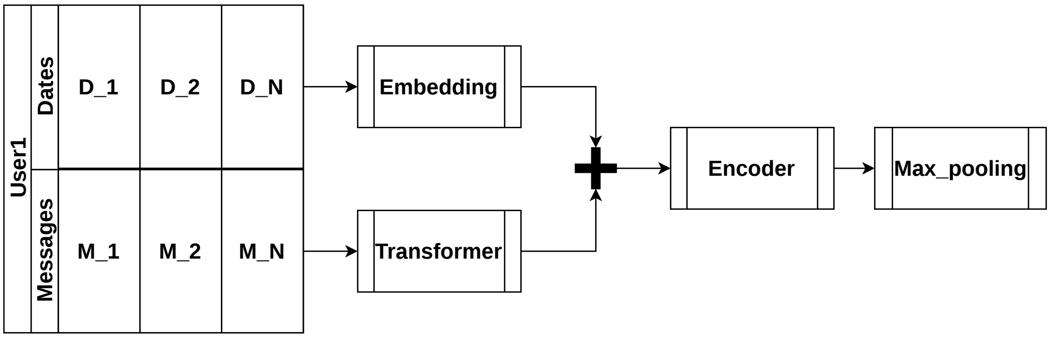

Figure 1: Architecture of the temporal hierarchical transformer.

through a dedicated embedding layer, designed to project chronological information into a continuous vector space. These temporal feature vectors are subsequently fused with the textual representations of the posts. This fusion mechanism enables the model to jointly learn content semantics and associated temporal patterns, thereby optimizing the identification of behaviors linked to circadian rhythms. The architecture of the temporal hierarchical transformer can be seen in Figure 1

3.2. Temporal Hierarchical Transformer time elapsed: Run 1

In this configuration, instead of encoding the specific hour and minute of each post, the system processes the time elapsed between consecutive messages. This approach enables the model to capture the user's posting dynamics and detect patterns such as bursts of activity or prolonged periods of inactivity, which may serve as significant behavioral markers within the context of depression. To process this sequence of intervals, a Long Short-Term Memory (LSTM) network is utilized to transform the elapsed-time tensor into a temporal embedding. This embedding is subsequently integrated with the textual representations through an element-wise addition. The resulting unified embedding effectively consolidates both the semantic content and the user's activity rhythm into a single feature space.

3.3. Ensemble majority voting: Run 2

The final system consists of a majority voting ensemble (hard voting). This aggregation strategy combines multiple base models by predicting the class that receives the highest number of votes among the constituent models. In our setup, an odd number of three models was selected to avoid tie-breaking constraints. The selected models correspond to the three configurations that achieved the highest performance during our local experimentation phase: the two Temporal Hierarchical Transformer variants, and one of the systems that was subsequently affected by the software bug during the official evaluation phase, which is based on hierarchical transformers.

4. Results

We now proceed to present and analyze the evaluation results for Task 2. For this challenge, the eRisk organizing committee established two different types of metrics. On one hand, classification performance is assessed using traditional computational indicators: Precision, Recall, and F1-score. On the other hand, the early detection nature inherent to this lab is quantified via the forum's standardized latency-based metrics: Early Risk Detection Error with cutoffs at 5 and 50 posts ($ERDE_5$ and $ERDE_{50}$), Latency-weighted True Positives ($Latency\ TP$), Speed and latency weighted F1 ($F_{latency}$).

In general terms, our framework achieved a highly competitive performance in this task, securing the second position in the global leaderboard and exhibiting narrow margins relative to the top-ranked team, as can be seen in Table 2. When analyzing the system behavior under early detection constraints, our approach excels by recording the most optimal performance in the $F_{latency}$ metric, sharing the first position along with another participating team. Additionally, the framework validates its temporal efficiency by consistently securing top-tier rankings across almost all other official latency metrics, thereby establishing an optimal equilibrium between classification accuracy and early clinical detection.

When contrasting the performance of run 0 against run 1, it becomes evident that the configuration explicitly incorporating discrete temporal features (hours and minutes) consistently outperforms the model that only computes linear elapsed time. These findings strongly converge with clinical and computational literature, which documents that individuals with depressive symptomatology exhibit severe disruptions in their circadian rhythms. This validates the predictive utility of publication hour coordinates as key behavioral biomarkers for the automated detection of depression.

Finally, the analysis of run 2, designed as a majority voting ensemble that integrates predictions from runs 0, 1, and one of the systems with the bug, reveals that its performance is severely bottlenecked by the individual limitations of the model affected by the bug. Upon evaluating the metrics, it becomes

Table 2
Task 2 decision-based evaluation is ordered in terms of best F1. The table is an edited version of the official one, where the runs with F1 lower than 0.65 have been removed. The complete table can be found in the overview paper

Team	Run	*P*	*R*	*F1*	*ERDE*	*ERDE*	lat_{TP}	*speed*	$F1_{lat}$
HUGETIME	0	0.80	0.87	**0.83**	0.15	**0.05**	9.00	0.97	**0.81**
UNED-GELP	0	0.79	0.85	0.82	0.11	0.06	4.00	0.99	**0.81**
	1	0.78	0.71	0.75	0.10	0.08	2.00	**1.00**	0.74
	2	0.81	0.69	0.75	0.11	0.08	3.00	0.99	0.74
NoirByte	0	0.88	0.76	0.82	0.15	0.09	20.00	0.93	0.76
	1	0.93	0.70	0.80	0.16	0.10	24.00	0.91	0.73
	2	0.93	0.62	0.74	0.16	0.11	28.50	0.89	0.66
	3	0.94	0.54	0.69	0.17	0.12	27.00	0.90	0.62
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
	1	0.69	0.89	0.78	0.15	**0.05**	9.00	0.97	0.75
	4	0.75	0.85	0.80	0.16	0.07	14.00	0.95	0.76
UET-PsyWar	0	0.62	0.87	0.72	0.15	0.08	15.00	0.95	0.68
	1	0.57	0.88	0.69	0.16	0.08	15.00	0.95	0.65
	2	0.73	0.82	0.77	0.15	0.07	15.00	0.95	0.73
	3	0.76	0.80	0.78	0.14	0.08	14.00	0.95	0.74
	4	0.60	0.88	0.71	0.15	0.07	15.00	0.95	0.68
LCDAA	0	0.66	0.81	0.73	0.11	0.07	3.00	0.99	0.72
	1	0.69	0.81	0.74	0.11	0.07	4.00	0.99	0.74
	2	0.78	0.78	0.78	0.13	0.08	10.00	0.96	0.75
	3	0.70	0.70	0.70	0.09	0.07	**1.00**	**1.00**	0.70
	4	0.72	0.74	0.73	0.11	0.07	2.00	**1.00**	0.73
Lotu-ixa	3	0.63	0.86	0.73	0.14	0.07	12.00	0.96	0.70
	4	0.75	0.77	0.76	0.12	0.07	6.50	0.98	0.74
erisk-cedri	0	0.73	0.77	0.75	0.10	0.07	1.50	**1.00**	0.75
VANGUARD	0	0.62	0.90	0.74	0.14	**0.05**	6.50	0.98	0.72
HUTECH-NLP	2	0.77	0.70	0.74	0.11	0.06	4.00	0.99	0.73
Snugi-AI-v2	0	0.72	0.71	0.72	0.18	0.08	8.00	0.97	0.70
	1	0.77	0.69	0.73	0.17	0.09	14.00	0.95	0.69
	2	0.71	0.66	0.69	0.18	0.07	8.00	0.97	0.67
	3	0.80	0.65	0.72	0.17	0.07	8.00	0.97	0.70
DeepCare	4	0.95	0.57	0.71	0.13	0.09	5.00	0.98	0.70

evident that run 0 fails to contribute synergistic value to the ensemble. Instead, the system consolidates a positive depression prediction primarily when convergence occurs between runs 0 and 1. This suggests that, given the neutralization of the first baseline, run 1 effectively operates as the critical deciding factor driving the final classification of the collective model.

5. Conclusions

In this work, we have presented a comprehensive methodological framework designed to address the challenges of continuous evaluation and early depression detection within social media environments. The official experimental results validate the competitiveness and robustness of our approaches, culminating in our system securing the second position in the global ranking.

The core contribution of this work is the Temporal Hierarchical Transformer, a novel architecture engineered to overcome the sequence length constraints inherent to standard Transformer variants. By enabling the simultaneous processing of long-form textual inputs and the seamless integration of explicit temporal information, this model has demonstrated substantial utility in handling the high density of user publications typical of online platforms. Ultimately, our findings prove that the proposed architecture is particularly effective in scenarios where chronological activity patterns serve as a highly discriminative feature for user classification, highlighting the critical value of time-aware modeling for automated early risk detection.

References

[1] W. H. Organization, et al., World mental health report: Transforming mental health for all (2022).

[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 (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.

[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, 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.

[4] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2018: Early risk prediction on the internet (extended lab overview), in: Proceedings of the 9th International Conference of the CLEF Association, CLEF, 2018, pp. 1–20.

[5] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.

[6] D. E. Losada, F. Crestani, J. Parapar, erisk 2020: Self-harm and depression challenges, in: European conference on information retrieval, Springer, 2020, pp. 557–563.

[7] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) 1 (2021) 864–887.

[8] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2022: pathological gambling, depression, and eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2022, pp. 436–442.

[9] 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.

[10] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2024: Depression, anorexia, and

eating disorder challenges, in: European Conference on Information Retrieval, Springer, 2024, pp. 474–481.

[11] 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.

[12] Q. Un Nisa, R. Muhammad, Towards transfer learning using bert for early detection of self-harm of social media users, Proceedings of the Working Notes of CLEF (2021) 21–24.

[13] A. M. Marmol-Romero, M. Garcia-Vega, M. A. Garcia-Cumbreras, A. Montejo-Raez, Sinai at erisk@ clef 2025: Transformer-based and conversational strategies for depression detection, arXiv preprint arXiv:2509.19861 (2025).

[14] R. Pan, J. A. García-Díaz, R. Valencia-García, Umuteam at erisk@ clef 2023 shared task: Transformer models for early detection of pathological gambling, depression, and eating disorder., in: CLEF (Working Notes), 2023, pp. 752–762.

[15] D. Guecha, A. Potdar, A. Miyaguchi, Ds@ gt erisk 2024: Sentence transformers for social media risk assessment, arXiv preprint arXiv:2407.08008 (2024).

[16] A. M. Mármol-Romero, F. M. Plaza-del Arco, A. Montejo-Ráez, Sinai at erisk@ clef 2023: Approaching early detection of gambling with natural language processing, arXiv preprint arXiv:2509.14797 (2025).

[17] 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.

[18] A. Wirz-Justice, Biological rhythm disturbances in mood disorders, International clinical psychopharmacology 21 (2006) S11–S15.

[19] N. Vedula, S. Parthasarathy, Emotional and linguistic cues of depression from social media, in: Proceedings of the 2017 International Conference on Digital Health, 2017, pp. 127–136.

[20] M. Ten Thij, K. Bathina, L. A. Rutter, L. Lorenzo-Luaces, I. A. van de Leemput, M. Scheffer, J. Bollen, Depression alters the circadian pattern of online activity, Scientific reports 10 (2020) 17272.

[21] D. Lim, J. Jeong, Y. M. Song, C.-H. Cho, J. W. Yeom, T. Lee, J.-B. Lee, H.-J. Lee, J. K. Kim, Accurately predicting mood episodes in mood disorder patients using wearable sleep and circadian rhythm features, NPJ Digital Medicine 7 (2024) 324.

[22] M. P. Lee, D. W. Kim, Y. Fang, R. Kim, A. S. Bohnert, S. Sen, D. B. Forger, The real-world association between digital markers of circadian disruption and mental health risks, NPJ digital medicine 7 (2024) 355.

[23] X. Dai, I. Chalkidis, S. Darkner, D. Elliott, Revisiting transformer-based models for long document classification, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 7212–7230.