XAI-Guided Transfer Learning: Supervised Attention for Contextualized Early Depression Detection

eRisk at CLEF 2026

Xabier Larrayoz[1,*], Alicia Pérez[1] and Panagiotis Papapetrou[2]

[1]HiTZ Center - Ixa (http://www.hitz.eus), University of the Basque Country (UPV/EHU), Spain
[2]Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

Abstract

This article describes the participation of the Lotu-ixa team in the eRisk 2026 Task 2: Contextualized Early Detection of Depression (2nd edition). We propose a *Supervised Attention* mechanism that transfers clinical knowledge extracted from historical eRisk datasets (2017–2022) into a hierarchical Transformer-based architecture designed to process full Reddit conversational threads. The core contribution is a *Dual Relevance Mask* built through Explainable AI (XAI) techniques—specifically SHAP attributions combined with lemmatisation and semantic soft-mapping over mental-health embeddings. This mask guides the model's internal attention toward diagnostically relevant tokens while suppressing conversational noise. We submitted five runs exploring different decision thresholds and a temporal risk-averaging strategy. Our best run achieves $F_1 = 0.76$ and $F_{lat} = 0.74$ with a median detection latency of 6.5 threads. Runs with lower thresholds exhibit very high Recall (≥ 0.93) at the cost of low Precision, a pattern that motivates the temporal aggregation strategy and points to future directions for calibration.

Keywords

early risk detection, explainable AI, supervised attention, transfer learning, eRisk CLEF

1. Introduction

The early detection of mental health disorders in social media has become an increasingly critical research challenge. The eRisk workshop [1] has been the principal benchmark for evaluating such systems under realistic, sequential conditions, where a system must process user writings incrementally and issue a risk alert as early as possible—the goal being to produce a correct classification while minimising the number of messages examined before the alert is triggered.

The 2026 edition of Task 2 marks the second iteration of the *Contextualized Early Detection of Depression* scenario [2, 3]. Unlike earlier editions that provided isolated posts, this task supplies entire Reddit discussion threads (including titles, third-party replies, and nested comments) forcing systems to reason about inter-participant dynamics in addition to the semantic nuances of the target user's language.

This shift introduces a fundamental challenge: **contextual noise**. In practice, more than 80% of the text in a typical Reddit thread is generated by participants other than the target user, covering topics entirely unrelated to that user's mental state. Standard Transformer models may suffer from *attention dispersion* [4], where clinically relevant signals authored by the target user are diluted by surrounding community text.

To address this, we propose an architecture based on **XAI-Guided Supervised Attention**. Instead of learning token relevance purely from the 2026 training data, our model is bootstrapped with an externally derived importance matrix $\mathbf{M}$, built from SHapley Additive exPlanations (SHAP) [5] attributions computed on historical eRisk data. This matrix acts as a spotlight: it assigns near-zero weights to noisy

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
xabier.larrayoz@ehu.eus (X. Larrayoz); alicia.perez@ehu.eus (A. Pérez); panagiotis@dsv.su.sel (P. Papapetrou)
0009-0003-7308-5965 (X. Larrayoz); 0000-0003-2638-9598 (A. Pérez); 0000-0002-4632-4815 (P. Papapetrou)

tokens and amplifies tokens that carry diagnostic value, effectively transferring clinical knowledge from isolated-post datasets to the noisier conversational setting.

We further propose a **Dual Relevance Mask** that encodes not only symptom-indicative vocabulary (henceforth Label 1) but also *protective factor* vocabulary (henceforth Label 0), forcing the model to consider evidence for both classes before committing to a decision. Five decision strategies are evaluated, ranging from aggressive early alerting to conservative threshold gating and historical risk averaging.

2. Related Work

Early risk detection in mental health has been explored extensively within the eRisk shared task series [6, 7, 8]. Prior work has addressed the challenge through lexicon-based approaches, sequential classifiers, and, more recently, Transformer-based models fine-tuned on mental-health corpora [9]. In contrast, our approach uses a general-purpose encoder [10] whose attention mechanism is guided by externally derived clinical knowledge, rather than relying on domain pre-training to implicitly encode mental-health vocabulary.

A persistent limitation in early detection of health risks in social media is that systems focus solely on isolated posts published by the target user, discarding the conversational context in which those posts were produced and, consequently, ignoring the interpersonal dynamics that can constitute the strongest evidence of risk [11]. The 2025 and 2026 editions of Task 2 explicitly target this gap by supplying full discussion threads, shifting the challenge from post-level classification to discriminating clinically relevant signals within noisy multi-participant exchanges.

On the explainability side, SHAP [5] has been applied to NLP models to identify feature importance at the token level. We build on this approach by using SHAP not merely as a post-hoc explanation tool, but as a *knowledge source* to supervise the attention mechanism during training—a direction explored in connection with attention regularisation [4].

3. Task Description and Data

3.1. Task Setup

Task 2 of eRisk 2026 is a sequential, interactive classification task. For each target user, the server releases one Reddit discussion thread per round. After processing the full thread (including every post by all interlocutors), teams must return a binary decision label (depressed / control) and a real-valued confidence score; upon submission, the next thread is unlocked.

Performance is measured using decision-based metrics (Precision, Recall, F_1, $ERDE_5$, $ERDE_{50}$) and the latency-weighted F-score $F_{latency}$ [12], as well as ranking-based metrics (P@10, NDCG@10, NDCG@100).

3.2. Datasets

Our approach draws on two complementary data sources, summarised in Table 1.

1. **Decontextualized source datasets (eRisk 2017–2022).** We use data from the eRisk 2017, 2018, and 2022 editions to establish a baseline for semantic importance. These datasets consist of isolated social media posts labelled at the user level, providing a clean signal for learning the linguistic markers of depression without conversational noise.
2. **Contextualized target dataset (eRisk 2026).** The training data for 2026 consists of the test set from the 2025 edition: full Reddit conversational threads from 909 users (102 depressed, 807 control). Each thread is structurally decomposed into three components:
    - *Start node*: the initial post that opens the thread (title + body).
    - *Target nodes*: all messages authored by the target user within the thread.
    - *Reply nodes*: all responses from other community members.

To maintain computational efficiency, each user's history is capped at 25 threads, retaining only the start node, target user messages, and direct replies to the target user.

Table 1
Summary of datasets. The 2026 eRisk edition introduces full conversational threads, whereas earlier editions provided decontextualized isolated posts.

eRisk edition	Contextualized	Users	Depressed	Control	Posts
2017 / 2018 / 2022	No	3,084	305	2,779	142,200
2026	Yes	909	102	807	104,916

4. Methodology

Our approach is divided into two sequential phases: (A) construction of the Dual Relevance Mask from the decontextualized source data, and (B) training of the hierarchical Transformer with supervised attention using the mask derived from the previous phase.

4.1. Phase A: Dual Relevance Mask Generation

The rationale of Phase A is to construct a Relevance Mask comprising two sets of tokens: those associated with alarm-related elements (henceforth referred to as *symptoms*) and those that discourage an alarm decision (henceforth referred to as *protective factors*). The procedure for constructing the mask is described below.

Step 1 – Base Model Training. A preliminary DistilBERT classifier [10] is fine-tuned on the decontextualized source datasets (2017, 2018, 2022) to learn depression markers without the distraction of conversational context.

Step 2 – XAI Attribution via SHAP. We apply SHAP (KernelExplainer) [5] to the trained base model to compute per-token attribution values. Tokens that consistently push the prediction toward the positive class (depression risk > 0.001, appearing in at least 3 distinct documents) are collected into the *symptom dictionary* $\mathcal{D}_+$. An analogous process on control-class texts produces the *protective-factor dictionary* $\mathcal{D}_-$. Both thresholds (risk > 0.001 and minimum document frequency of 3) are empirical design choices whose sensitivity is discussed in Section 5.4.

Step 3 – Linguistic Generalisation. All tokens are lemmatised using spaCy [13] to ensure morphological coverage. Out-of-vocabulary terms found in the 2026 dataset are handled through semantic soft-mapping: cosine similarity over DistilBERT contextual embeddings transfers the importance weight of known terms to unseen synonyms (e.g., the weight of *sad* may be transferred to *melancholy*).

Step 4 – Purification and Fusion. Ambiguous terms appearing in both $\mathcal{D}_+$ and $\mathcal{D}_-$ are removed:

$$\begin{aligned} \mathcal{D}_\cdot &\leftarrow \mathcal{D}_+ \cap \mathcal{D}_- \\ \mathcal{D}_+ &\leftarrow \mathcal{D}_+ \setminus \mathcal{D}_\cdot \\ \mathcal{D}_- &\leftarrow \mathcal{D}_- \setminus \mathcal{D}_\cdot \end{aligned}$$

The purified dictionaries are then combined into the Dual Relevance Mask $\mathbf{M}$, formally defined per token t as:

$$\mathbf{M}[t] = \begin{cases} +1.0, & \text{if } t \in \mathcal{D}_+ \\ -1.0, & \text{if } t \in \mathcal{D}_- \\ 0.0, & \text{if } t \notin \mathcal{D}_+ \cup \mathcal{D}_- \end{cases} \tag{1}$$

By assigning a non-zero weight to both symptom and protective-factor tokens, the model is forced to read evidence from both classes before the final classification layer makes a context-aware decision, deliberately avoiding *attentional blindness*.

4.2. Phase B: Supervised Attention Architecture

4.2.1. Standard Self-Attention Baseline

In standard self-attention [14], token relevance is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q}$ (queries), $\mathbf{K}$ (keys), and $\mathbf{V}$ (values) are learned linear projections of the input, and d_k is the dimension of the key vectors, which serves as a scaling factor to stabilise gradients.

4.2.2. Diagnostic Bias Matrix

We introduce a bias matrix $\mathbf{M}$, derived from the Dual Relevance Mask (1), that modifies the attention scoring:

$$\text{Attention}_{\text{supervised}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \tag{3}$$

For a given token t, the bias value $\mathbf{M}[t]$ is set according to (1): tokens in $\mathcal{D}_+$ or $\mathcal{D}_-$ receive a non-zero weight that amplifies their signal, while tokens in neither dictionary (i.e. conversational noise) receive 0.0 and are thus left unmodified by the bias.

4.2.3. Hierarchical Risk Aggregation

Each thread is decomposed into the three channels introduced in Section 3: start node, target nodes, and reply nodes. Both architectures process these channels independently using DistilBERT with chunking (128 tokens per chunk, up to 100 chunks per channel). Each channel produces a risk score $r \in [0, 1]$ from the final classification head of the encoder. The channel-level scores are then aggregated as:

$$R_{\text{thread}} = w_1 \cdot r_{\text{start}} + w_2 \cdot r_{\text{target}} + w_3 \cdot r_{\text{reply}} \tag{4}$$

where w_1, w_2, w_3 are learned via a softmax over trainable parameters.

4.2.4. Architectures

In our experiments two architectures were explored:

- **Baseline Risk Aggregator**: standard hierarchical classifier without attention supervision, as in (2).
- **Contextual Risk Aggregator**: identical architecture augmented with the supervised attention layer, as in (3), where $\mathbf{M}$ is generated in real time during tokenisation from the Dual Relevance Mask.

4.2.5. Dual Loss Training

Training optimises a joint objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{XAI}} \tag{5}$$

where $\mathcal{L}_{\text{CE}}$ is standard cross-entropy for the classification task, $\mathcal{L}_{\text{XAI}}$ is the mean squared error between the model's internal attention heads and the pre-computed SHAP relevance masks, and λ is a regularisation hyperparameter balancing the two objectives.

4.3. Decision Strategy

Once R_{thread} is computed for a given user's current thread, a binary alert is issued according to a decision threshold τ: the user is flagged as at risk if $R_{\text{thread}} \geq \tau$. In addition to single-thread thresholding, we explore a *temporal memory* strategy in which the alert is triggered when the rolling mean of the three most recent thread-level scores exceeds τ, absorbing isolated spikes and targeting users with persistent rather than transient distress signals. The five concrete configurations submitted to eRisk are described in Section 5.1.

4.4. Training Protocol

To prevent data leakage in the presence of class imbalance (102 depressed vs. 807 control users), we apply **5-fold stratified cross-validation at the user level**, ensuring that no fragment of any user's messages is shared between training and test folds. The base model is DistilBERT (`distilbert-base-uncased`), pre-trained then fine-tuned under this protocol.

5. Experimental Framework

The Lotu-ixa team submitted five runs varying several aspects explored in the methodology; these runs are detailed in Section 5.1. The official eRisk 2026 preliminary results are described next: Sections 5.2 and 5.3 are devoted to the decision-based and ranking-based results, respectively, and Section 5.4 discusses the experimental findings.

5.1. Experimental Setup

For the official eRisk submission, we designed a streaming client that processes sequential rounds in real time. At each round, the client receives a new discussion thread for every target user, runs inference with both architectures, and must return a binary decision and a confidence score before the next thread is released.

Rather than submitting a single system, we exploited the five allowed runs to empirically explore the Precision–Recall trade-off under different operating conditions. This design serves a twofold purpose: (i) it provides a systematic ablation of the threshold sensitivity of the Contextual Risk Aggregator, and (ii) it allows direct comparison between the XAI-guided model and the unguided baseline at the same decision threshold. Table 2 summarises the five submitted configurations.

Table 2
Five approaches submitted by the Lotu-ixa team. BRA: Baseline Risk Aggregator (standard self-attention, Eq. 2); CRA: Contextual Risk Aggregator (supervised attention, Eq. 3). The decision strategy column indicates whether a fixed threshold or a rolling mean (see Section 4.3) is applied.

Run	Arch.	τ	Decision strategy
0	BRA	0.50	**Unguided baseline.** No supervised attention. Isolates the contribution of the XAI layer.
1	CRA	0.50	**Aggressive early alert.** Maximises Recall; accepts a higher false-positive rate to avoid missing any true case.
2	CRA	0.70	**Balanced.** Higher threshold to filter false alarms from control users who occasionally post on negative topics.
3	CRA	0.85	**Conservative.** Alert only when symptoms are strongly present and protective factors absent. Maximises Precision at the cost of Recall and latency.
4	CRA	0.60	**Temporal memory.** Rolling mean of the three most recent thread-level risk scores (Section 4.3). Fires if the mean exceeds τ, absorbing single-episode spikes and targeting chronically at-risk users.

5.2. Decision-Based Results

Table 3 shows the full competition ranking ordered by best F_1 per team, with Lotu-ixa's five runs highlighted. The top-performing systems (HUGETIME, UNED-GELP, NoirByte) achieve $F_1 \geq 0.82$ with latencies between 4 and 20 threads; our best run (Run 4) achieves $F_1 = 0.76$ at a latency of 6.5 threads, placing it in the upper-middle tier of the competition.

Table 3
Decision-based evaluation results (Task 2, eRisk 2026), ordered by best F_1. Lotu-ixa runs are highlighted in bold.

Team	Run	P	R	F_1	$ERDE_5$	$ERDE_{50}$	lat_{TP}	F_{lat}
HUGETIME	0	0.80	0.87	0.83	0.15	0.05	9.0	0.81
UNED-GELP	0	0.79	0.85	0.82	0.11	0.06	4.0	0.81
	1	0.78	0.71	0.75	0.10	0.08	2.0	0.74
	2	0.81	0.69	0.75	0.11	0.08	3.0	0.74
NoirByte	0	0.88	0.76	0.82	0.15	0.09	20.0	0.76
MindAILab	0	0.84	0.80	0.82	0.16	0.08	16.0	0.77
INSA-Lyon	0	0.76	0.85	0.80	0.16	0.07	13.0	0.76
DeepCare	4	0.95	0.57	0.71	0.13	0.09	5.0	0.70
LCDAA	2	0.78	0.78	0.78	0.13	0.08	10.0	0.75
UET-PsyWar	3	0.76	0.80	0.78	0.14	0.08	14.0	0.74
Lotu-ixa	**4**	**0.75**	**0.77**	**0.76**	**0.12**	**0.07**	**6.5**	**0.74**
	3	0.63	0.86	0.73	0.14	0.07	12.0	0.70
	0	0.37	0.96	0.54	0.13	0.07	3.0	0.53
	2	0.38	0.93	0.54	0.14	0.08	6.0	0.53
	1	0.31	0.97	0.47	0.14	0.08	3.0	0.46
erisk-cedri	0	0.73	0.77	0.75	0.10	0.07	1.5	0.75
VANGUARD	0	0.62	0.90	0.74	0.14	0.05	6.5	0.72
HUTECH-NLP	2	0.77	0.70	0.74	0.11	0.06	4.0	0.73
Snugi-AI-v2	1	0.77	0.69	0.73	0.17	0.09	14.0	0.69
NYCU-NLP	3	0.49	0.64	0.56	0.17	0.11	15.0	0.52
Awakened	0	0.33	0.68	0.45	0.11	0.10	1.0	0.45
MacEarlyDetect	2	0.22	0.98	0.36	0.17	0.11	3.0	0.36

5.3. Ranking-Based Results

Beyond binary decisions, eRisk also evaluates the quality of continuous risk scores as a ranking instrument via NDCG@100. Table 4 reports NDCG@100 at four writing checkpoints for the top systems and all five Lotu-ixa runs. A well-calibrated system should see NDCG@100 *increase* as more context accumulates; systems where it decreases reveal a score drift problem.

5.4. Discussion

The results reveal a clear and instructive pattern across all dimensions of evaluation.

Run 4 is the overall best system. The rolling mean strategy (Run 4, $\tau = 0.60$) achieves $F_1 = 0.76$ and $F_{lat} = 0.74$, with a median detection latency of 6.5 threads. This places it on par with several top-ranked teams such as erisk-cedri ($F_1 = 0.75$, $F_{lat} = 0.75$) and Lotu-ixa's own Run 3 ($F_1 = 0.73$). The result validates the core motivation of temporal aggregation: by computing a rolling mean over the last three threads rather than reacting to a single observation, the system suppresses isolated spikes and focuses on users who exhibit *persistent* rather than transient distress signals.

Runs 0–2 expose a precision collapse under low thresholds. Runs 0, 1, and 2 exhibit very high Recall (0.93–0.97) but extremely low Precision (0.31–0.38), yielding F_1 scores of 0.47–0.54. This

Table 4
Ranking-based evaluation: NDCG@k at selected writing checkpoints. Top systems (best run only) and all Lotu-ixa runs are shown. Column headers indicate the number of writings observed.

Team	Run	@1	@100	@250	@500
DeepCare	0	0.74	0.88	0.90	0.91
MindAILab	0	0.69	0.86	0.89	0.90
NoirByte	0	0.72	0.86	0.88	0.89
INSA-Lyon	0	0.64	0.84	0.87	0.89
LCDAA	0	0.61	0.84	0.87	0.88
HUTECH-NLP	0	0.71	0.82	0.84	0.85
UNED-GELP	0	0.68	0.87	0.82	0.83
erisk-cedri	0	0.65	0.81	0.84	0.84
Lotu-ixa	0	0.63	0.56	0.41	0.23
	1	0.60	0.59	0.42	0.25
	2	0.60	0.59	0.42	0.25
	3	0.60	0.59	0.42	0.25
	4	0.67	0.58	0.49	0.38

imbalance means that for every true positive alert, the system issues roughly two to three false positives. Two factors likely contribute to this behaviour. First, the class imbalance in the training data (102 depressed vs. 807 control, a ratio of approximately 1:8) biases the model toward over-predicting the positive class. Second, at thresholds $\tau \leq 0.70$, the Contextual Risk Aggregator appears insufficiently calibrated: the XAI-guided attention amplifies depressive signals effectively, but the final risk score is not well-separated between true positives and hard negatives (control users who also post emotionally negative content). Notably, the XAI model (Run 1) does *not* outperform the baseline (Run 0) in F_1 at the same threshold, suggesting that at $\tau = 0.50$ the additional attention supervision has a negligible effect on the decision boundary and that calibration is the dominant bottleneck.

Threshold gating (Run 3) recovers discrimination. Raising the threshold to $\tau = 0.85$ (Run 3) substantially improves Precision to 0.63 while maintaining Recall of 0.86, yielding $F_1 = 0.73$. This demonstrates that the underlying risk scores do contain the signal necessary to separate the classes—the scores are simply not well-calibrated at lower operating points. The increased latency (12.0 threads) is the cost of this conservatism: the model requires more evidence before crossing such a high threshold.

Ranking performance degrades over time. Table 4 reveals a striking pattern: Lotu-ixa achieves strong early-stage ranking (NDCG@100 ≈ 0.60 at 1 writing for most runs) but this degrades sharply as more context is accumulated, falling to NDCG@100 $= 0.23$–0.38 at 500 writings. This is in stark contrast to top systems (DeepCare, NoirByte), whose NDCG@100 improves monotonically with more evidence. We interpret this as a *score drift* problem: as additional threads are processed, the rolling risk scores for both positive and negative users converge toward similar values, reducing the discriminative power of the ranking. A likely cause is that the weighted aggregation in Equation (4) averages out the risk signal over time for users with heterogeneous posting behaviour, rather than retaining the peak or cumulative maximum signal. This is a known limitation of mean-based aggregation in streaming settings and motivates future exploration of alternatives such as exponential decay weighting or maximum-risk-over-history strategies.

ERDE scores and early detection speed. $ERDE_5$ and $ERDE_{50}$ values are consistent across all our runs (0.12–0.14 and 0.07–0.08 respectively), and competitive with mid-tier systems in the competition. The speed factor F_{lat} is highest for Run 4 (0.74), confirming that the rolling mean does not critically delay true positive detections while improving Precision.

6. Conclusions

To tackle the contextualised early detection of depression, we have presented a Supervised Attention approach that injects clinical knowledge directly into the attention mechanism of a DistilBERT-based hierarchical classifier. The clinical knowledge injected is derived from SHAP attributions computed over an earlier collection of decontextualized eRisk data (2017–2022).

The key contributions are: (i) a Dual Relevance Mask (Section 4.1) that encodes both symptom and protective-factor vocabularies at equal positive weight, deliberately avoiding attentional blindness; (ii) a hierarchical three-channel architecture that separately models the start node, target-user messages, and community replies; and (iii) a dual loss combining cross-entropy with an attention-alignment term.

Results show that the temporal aggregation strategy (Run 4) is the most effective decision policy, achieving $F_1 = 0.76$ and $F_{lat} = 0.74$, competitive with several top-ranked teams in the competition. However, runs operating at lower thresholds ($\tau \leq 0.70$) exhibit a precision collapse that reveals two open problems: (a) the model is not well calibrated under class imbalance, and (b) the rolling-mean aggregation causes score drift that degrades ranking quality as more context is accumulated.

Future work should address these limitations, possibly through: (a) post-hoc calibration (e.g., temperature scaling or Platt scaling) to improve the operating-point independence of the risk score; (b) replacing mean-based aggregation with maximum-risk or exponential-decay weighting to preserve the ranking signal over longer histories; (c) substituting DistilBERT with a domain-adapted encoder (MentalBERT or similar) as the base encoder; and (d) extending the Dual Relevance Mask with ontology-derived clinical constructs aligned with DSM-5 symptom criteria.

Acknowledgments

This work has been partially supported by the HiTZ Center and the following MCIN/AEI/10.13039/501100011033 project: EDHIA (PID2022-136522OB-C22). The first author is recipient of a grant from the Spanish Ministry of Education's *Formación de Profesorado Universitario* program (FPU23/01068).

Declaration on Generative AI

During the preparation of this work, the author(s) used generative AI tools exclusively for grammar and spelling checking. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF), Springer, 2016, pp. 28–39.

[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] S. Wiegreffe, Y. Pinter, Attention is not explanation, arXiv preprint arXiv:1908.04626 (2019).

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), volume 30, 2017.
[6] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the Internet: Experimental foundations, in: Proceedings of CLEF 2017, Springer, 2017.
[7] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early risk prediction on the Internet, in: Proceedings of CLEF 2018, Springer, 2018.
[8] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2022: Early risk prediction on the Internet, in: Proceedings of CLEF 2022, Springer, 2022.
[9] 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 Language Resources and Evaluation Conference (LREC), 2022, pp. 7184–7190.
[10] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.
[11] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2025: Early risk prediction on the Internet, in: Proceedings of CLEF 2025, Springer, 2025.
[12] F. Sadeque, D. Xu, S. Bethard, UArizona at the CLEF eRisk 2018 tasks: Precision medicine and early detection of mental health disorders, in: Proceedings of CLEF 2018, Springer, 2018.
[13] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). doi:10.5281/zenodo.1212303.
[14] P. Shaw, J. Uszkoreit, A. Vaswani, Self-attention with relative position representations, CoRR abs/1803.02155 (2018). URL: http://arxiv.org/abs/1803.02155. arXiv:1803.02155.