Leveraging Conversational Context and Semantic Relabeling for Early Depression Detection

Notebook for the eRisk Lab at CLEF 2025

Xabier Larrayoz*, Arantza Casillas and Alicia Pérez

HiTZ Center - Ixa (http:// www.hitz.eus), University of the Basque Country (UPV/EHU), Spain

Abstract

Early detection of depression in online interactions is critical for timely intervention and support. The aim of this work is to identify users at risk in full conversational threads by contrast to previous works approaching the task in isolated posts from a single user. Our approach comprises three key components: (1) a semi-supervised semantic relabeling of training data using transformer-based embeddings and percentile-based score thresholds to reduce label noise; (2) fine-tuning of a multilingual transformer model for binary depression risk classification; and (3) an inference pipeline that computes per-thread user and context risk scores and fuses them into a global, cumulative risk measure. Our method was evaluated under both decision-based and ranking-based paradigms. While a conservative decision threshold limited precision in the classification task, our system achieved top-tier performance in ranking-based metrics (Precision and NDCG), demonstrating the efficacy of contextual signal fusion for early depression detection. We discuss the impact of fusion weights on early detection error (ERDE), latency, and overall F_1, and outline future directions involving adaptive thresholding and advanced context encoding.

Keywords

early detection of depresion, social media, generative large language models, natural language understanding

1. Introduction

Early detection of signs of depression on digital platforms has gained increasing relevance in medical informatics given its potential to contribute to preventive mental health interventions. Related tasks focused on the analysis of isolated user messages, without considering the interactional context surrounding each post. In this work a more realistic scenario is tackled as complete conversation threads are available. This enables the exploitation of dialogical dynamics to identify clinically relevant indicators that emerge only within the framework of exchanges among participants. Thus, the goal here is to cope with Contextualized Early Detection of Depression.

Our approach is based on three main components: (i) a semantic relabeling of the training data to mitigate the noise in the original labels; (ii) the adaptation and fine-tuning of a multilingual transformer model for binary classification of depression risk; and (iii) an inference strategy that combines individual and contextual risk signals in real time. The results show that, although the decision threshold used limited performance in the decision-based evaluation, our methodology achieved highly competitive results in the ranking-based evaluation, confirming the effectiveness of incorporating conversational context in the early detection of depression.

The structure of this paper is as follows: Section 2 reviews related work; Section 3 presents the task definition and the available data; Section 4 describes the proposed system; Section 5 discusses the results; and finally, Section 6 outlines the conclusions and future research directions.

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

*Corresponding author.

†These authors contributed equally.

xabier.larrayoz@ehu.eus (X. Larrayoz); arantza.casillas@ehu.eus (A. Casillas); alicia.perez@ehu.eus (A. Pérez)

0009-0003-7308-5965 (X. Larrayoz); 0000-0003-4248-8182 (A. Casillas); 0000-0003-2638-9598 (A. Pérez)

2. Related Work

Early detection of mental health disorders has been a central theme in all editions of eRisk [1, 2, 3], becoming a benchmark task in the field of longitudinal data analysis in digital environments. Over the years, the early detection task has focused on various clinical conditions: in 2022 the task addressed depression [1]; in 2023 pathological gambling [2]; and in 2024 signs of anorexia [3]. In all cases, the main objective has been to issue an early alert about potential risk cases based on users' posts in online forums.

In the 2024 edition, for instance, the NLP-UNED team proposed a system for the early detection of signs of anorexia that included several key components [4]. Their approach began with a semantic representation of messages using sentence encoders, followed by a relabelling process based on Approximate Nearest Neighbors (ANN) techniques, allowing them to transform a dataset originally annotated at the user level into one labelled at the message level. They further refined the embeddings using contrastive learning, aiming to maximize the distance between examples from different classes. For classification, they also relied on ANN methods combined with rules and heuristics to expand the number of messages considered per user for each prediction. Their system achieved the best results in both the decision-based and the ranking-based evaluations.

Noteworthy, Riewe-Perła and Filipowska [5] combined language models with recommender systems to predict whether recommended content originated from individuals with mental health conditions. They used document and user embeddings along with a hybrid recommendation engine (LightFM [6]), built on sentence transformers (SBERT [7]), to efficiently and early classify signals of potential risk.

Beyond eRisk, the MentalRiskES competition series has also addressed early detection tasks across different disorders, including depression, anxiety, and gambling [8, 9]. However, in all of these cases, the focus has remained on analyzing isolated user messages, without incorporating the conversational context in which these messages appear.

In contrast, the 2025 edition of eRisk [10, 11] introduces a paradigm shift by providing complete conversational threads, allowing for the study of interactive dynamics between participants. This new perspective enables a more realistic and clinically relevant analysis framework, in which risk indicators may arise not only from the target user's messages but also from the surrounding conversational context.

The availability of context brings clear advantages, such as enabling co-reference resolution and a better understanding of discourse flow, which are essential for detecting subtle signals of distress. However, it also introduces challenges: longer inputs increase computational cost and latency, and models must carefully focus on the target user's contributions without being distracted by irrelevant surrounding content. This setting requires new modeling strategies that balance depth of analysis with early response efficiency, opening both opportunities and challenges for future work.

3. Task Definition and Data Description

In this work we are dealing with Task 2 of eRisk 2025, entitled *Contextualized Early Detection of Depression*. This task introduces a novel scenario in depression detection as it leverages full conversational context. Unlike previous editions of eRisk, where only isolated messages authored by each user were released, the 2025 task provides participants with complete Reddit discussion threads involving the target user. Thus, during the test phase, systems receive not only the messages written by the user under analysis but also every other contribution within the thread, including the interaction structure that links the posts. This design stems from the observation that the clinical relevance of a message often becomes apparent only when interpreted in light of the surrounding conversation—for instance, an apparently neutral reply may reveal signs of hopelessness when responding to criticism or a plea for help. The task thereby simulates real-world scenarios in which detecting depression requires analyzing exchanges between multiple participants.

During the **training phase**, participants worked with a static corpus of 3,084 users derived from

earlier editions of eRisk (2017, 2018, and 2022), which included only the messages written by the target user, with no conversational context. This dataset contains an average of 640 messages per user, with 2,772 users labeled as control (label 0) and 312 as depressed (label 1). The distribution across editions is as follows: 1,400 users from the 2022 edition, 864 from 2017, and 820 from 2018. While the aim is to detect the risk for a user with the user in a conversation with other users, the training corpus lacks of contextual information, the entire thread is unavailable and this represents a challenge to train and adapt the models to a realistic situation. To mitigate this mismatch, we explored strategies to reinterpret the original annotations through a message-level semantic relabeling, and to design a decision-making mechanism capable of aggregating risk signals over time in a way compatible with the interactive structure of the test data.

In the **test phase**, the eRisk 2025 server operated in an interactive manner: for each target user, a new discussion thread was released in real time. Each thread constituted a submission round and included all posts published up to that point, including those written by other interlocutors. After processing the thread—taking into account both the content and its conversational structure—participants had to return a binary label (positive/negative) along with a confidence score before the next thread became available. This incremental protocol emulates a continuous monitoring environment in which access to the full conversational context is crucial for early and effective risk assessment.

To evaluate system performance, two complementary paradigms are used. First, the **decision-based evaluation** focuses on the final label assigned to each user and the moment at which a positive prediction is made. In this setting, classical classification metrics—precision, recall, and F1—are calculated, along with the ERDE (Early Risk Detection Error) metric [12], which penalizes both late detections of true positive cases and premature false positives. Additionally, the average detection latency (defined as the average number of threads processed before the first positive prediction) is computed and combined with the F1-score to obtain a latency-weighted F1, thus, balancing accuracy and promptness. Second, the **ranking-based evaluation** uses the confidence scores after a fixed number of rounds (e.g., after 1, 100, or 500 threads) to rank users by their estimated risk. Metrics such as Precision@K and NDCG@K are applied to assess the ability of the system to prioritize true positive cases. Together, these evaluation paradigms offer a comprehensive view of system performance, both in terms of classification quality and timeliness of detection.

4. Proposed Methodology

Our approach is structured into three main components, respectively explored in subsequent sections: a semantic relabeling process applied to the training data, the design and fine-tuning of a classification model, and an inference strategy based on the fusion of risk signals derived from both the target user and the surrounding conversational context.

4.1. Training set semantic re-labeling

Although the final objective is to assign a risk label at the user level, our model operates at the message level, processing each post independently. A common approach is to propagate the user's label to all their messages, treating every post from a depressed user as a positive instance. However, this assumption introduces a significant amount of noise, since not all messages from a depressed user necessarily exhibit linguistic markers of depression. Training under such noisy supervision can lead the model to learn spurious correlations or to dilute the signal of truly informative posts. To address this challenge, we designed a message-level relabeling strategy aimed at enhancing label fidelity and improving the classifier's robustness.

In order to reduce noise in the original labels and improve training quality, we implemented a semi-supervised message-level relabeling process. This methodology, inspired by prior work from the NLP-UNED group [4], as well as other related approaches proposed by different studies [13, 14]. Adapting this strategy to our specific task, we leveraged semantic representations obtained through pre-trained embeddings. For each message, a similarity score was computed with respect to representative

positive and negative examples, and a percentile-based strategy was applied to determine which messages were suitable for relabeling. This process allowed for precise control over the proportion of modified instances, preserving semantic consistency across the corpus while mitigating the effects of incorrect labels.

4.2. Classifier

The classifier is based on a multilingual transformer architecture with 12 layers and an embedding dimension of 768, derived from the XLM-RoBERTa model family [15]. The model was fine-tuned on the relabeled dataset (mentioned in section 4.1) using a binary cross-entropy loss function. Several combinations of hyperparameters (learning rate, batch size, and number of epochs) were explored through grid search, and the best configuration was selected based on performance over a validation set.

The model operates at the message level: each post is processed independently, and the classifier outputs a probability score indicating the likelihood that the message reflects signs of depression. These individual message-level scores do not directly determine the user's risk label; instead, they serve as input for a subsequent aggregation phase. In the next subsection, we describe how these scores are combined at the thread level to compute a global risk score per user, which is ultimately used for the final decision-making.

4.3. Decision making via Risk Signal Fusion

During the test phase, the system receives, sequentially, complete conversation threads in which the target user has participated. For each thread received at time t, three risk scores are computed:

- **User risk score** $r_u(t)$: computed as the mean depression probability of all messages written by the target user within the thread.
- **Context risk score** $r_c(t)$: computed as the mean depression probability of all other messages in the thread not authored by the target user.
- **Thread risk score** $r(t)$: We explored three alternatives:
  - Weighted linear combination of user risk and context risk, denoted as r_{WL} and computed as in expression (1), where the parameter $\alpha \in [0, 1]$ controls the relative influence of the user versus the context. This parameter, α, was empirically tuned during validation.
  - Mean message risk r_{Av}: Thread risk score computed as the mean depression probability of all messages in the thread regardless of author as in (2).
  - Maximum: Thread risk score equal to the maximum between user risk and context risk denoted as r_{Max} in (2).

$$r_{WL}(t) = \alpha \cdot r_u(t) + (1 - \alpha) \cdot r_c(t) \tag{1}$$

$$r_{Av}(t) = \frac{1}{|M_t|} \sum_{m \in M_t} r(m) \tag{2}$$

$$r_{Max}(t) = \max_{m \in M_t} r(m) \tag{3}$$

Based on the thread-level scores observed up to time t, a *global risk score* is calculated as the average of all previous thread scores, as in (4).

$$R_{\text{global}}(t) = \frac{1}{t} \sum_{i=1}^{t} r(i). \tag{4}$$

Finally, the binary decision regarding whether the user is at risk of depression is obtained by comparing $R_{\text{global}}(t)$ to a threshold τ, which was also optimized on the validation set. This strategy enables the system to incrementally integrate both the behavioral progression of the target user and the

evolving conversational dynamics. Moreover, for the ranking-based evaluation, the confidence score assigned to each user corresponds directly to the value of $R_{\text{global}}(t)$, thus preserving the relative risk intensity inferred from the conversation.

5. Results and Discussion

In our experiments, we conducted five configurations (runs) distinguished by how the thread risk score is computed, as presented in section 4.3. Table 1 summarizes the configurations by run.

Table 1
Summary of run configurations.

Run	r	α	Description
R0	r_{WL}	1.0	User risk only
R1	r_{WL}	0.8	80% user + 20% context
R2	r_{WL}	0.6	60% user + 40% context
R3	r_{Av}	–	Mean of probabilities of all thread messages
R4	r_{Max}	–	Maximum of user and context risk

Table 2 shows the best runs of each team under the decision-based evaluation paradigm, including all five configurations of our approach, denoted as Lotu-Ixa. We report precision (P), recall (R), F_1, $ERDE_5$, $ERDE_{50}$, true positive latency ($latency_{TP}$), speed, and latency-weighted F_1 ($F_{latency}$).

Table 2
Decision-based evaluation for Task 2.

Team	Run	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	$F_{latency}$
HIT-SCIR	4	0.77	0.94	**0.85**	0.09	**0.03**	8.0	0.97	**0.82**
ELiRF-UPV	0	**0.78**	0.81	0.79	0.08	0.04	7.0	0.98	0.78
HU	1	0.72	0.77	0.75	0.10	0.05	11.0	0.96	0.72
UET-Psyche-Warriors	2	0.63	0.86	0.73	0.09	0.04	16.0	0.94	0.68
PJs-team	0	0.66	0.75	0.71	0.09	0.06	17.0	0.94	0.66
Lotu-Ixa	0	0.43	0.79	0.56	**0.05**	0.04	2.0	**1.00**	0.56
	1	0.46	0.79	0.58	**0.05**	**0.03**	2.0	**1.00**	0.58
	2	0.47	0.79	0.59	**0.05**	**0.03**	2.0	**1.00**	0.59
	3	0.53	0.78	0.63	**0.05**	**0.03**	**1.0**	**1.00**	0.63
	4	0.15	**1.00**	0.25	0.09	0.08	**1.0**	**1.00**	0.25
COTECMAR-UTB	0	0.29	0.65	0.40	0.12	0.10	69.0	0.74	0.29
SINAI-UJA	0	0.24	**1.00**	0.39	0.08	0.05	3.0	0.99	0.38
NYCUNLP	4	0.20	0.93	0.33	0.16	0.07	18.0	0.93	0.31
FU-TU-DFKI	0	0.17	0.97	0.29	0.16	0.07	11.0	0.96	0.28
Capy-team	0	0.11	**1.00**	0.20	0.11	0.10	1.5	**1.00**	0.20
DS-GT	0	0.11	**1.00**	0.20	0.12	0.10	2.0	**1.00**	0.20

Our team achieved the best performance in $ERDE_5$, $ERDE_{50}$, $latency_{TP}$, and speed, demonstrating the effectiveness of our early detection strategy. Of our five configurations, R3 (mean probability of all thread messages) achieved the best balance, reaching $ERDE_5 = 0.05$, $ERDE_{50} = 0.03$, the lowest true-positive latency (1 thread), and optimal speed (1.00). Configuration R4 (max-risk fusion) performed worst, confirming that emphasizing only the highest-risk message is suboptimal. Additionally, we observe progressive improvements in ERDE and latency as the context weight increases up to R3, validating the usefulness of conversational context.

Table 3 presents ranking-based metrics for each team's best run, including the five configurations submitted by our team. Evaluated at various writing thresholds (1, 100, 500, and 1000), our approach demonstrates highly competitive performance in Precision and NDCG across all cutoffs, consistently ranking among the top participants.

Our approach has been competitive across all ranking metrics, ranking among the top participants. This supports our motivation to opt for risk signal fusion approach and we feel that further efforts

Table 3
Confidence ranking-based evaluation for Task 2 (best run per team compared to all runs by Lotu-Ixa).

Team	Run	1 writing			100 writings			500 writings			1000 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
HIT-SCIR	0	**1.00**	**1.00**	0.58	**1.00**	**1.00**	**0.84**	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
ELiRF-UPV	0	0.90	0.88	0.36	**1.00**	**1.00**	0.69	0.90	0.94	0.74	0.90	0.81	0.74
HU	1	**1.00**	**1.00**	**0.62**	0.90	0.88	0.57	0.60	0.71	0.35	0.40	0.60	0.26
UET-Psyche-Warriors	1	0.90	0.93	0.43	0.10	0.12	0.11	0.00	0.00	0.12	0.00	0.00	0.12
PJs-team	0	0.60	0.59	0.35	0.50	0.44	0.38	0.70	0.78	0.60	0.60	0.69	0.63
Lotu-Ixa	0	0.80	0.84	0.55	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.62	**1.00**	**1.00**	0.64
	1	0.90	0.94	0.57	**1.00**	**1.00**	0.73	**1.00**	**1.00**	0.63	**1.00**	**1.00**	0.63
	2	**1.00**	**1.00**	0.58	**1.00**	**1.00**	0.73	**1.00**	**1.00**	0.61	**1.00**	**1.00**	0.62
	3	0.90	0.81	0.58	**1.00**	**1.00**	0.74	**1.00**	**1.00**	0.61	**1.00**	**1.00**	0.62
	4	0.60	0.59	0.44	0.80	0.84	0.55	0.90	0.94	0.52	**1.00**	**1.00**	0.52
COTECMAR-UTB	0	0.30	0.23	0.23	0.00	0.00	0.22	0.20	0.15	0.18	0.20	0.13	0.17
SINAI-UJA	0	**1.00**	**1.00**	0.59	0.80	0.87	0.53	0.90	0.88	0.54	0.90	0.92	0.54
NYCUNLP	0	0.50	0.53	0.42	0.70	0.68	0.35	0.70	0.62	0.33	0.50	0.47	0.31
FU-TU-DFKI	0	0.90	0.94	0.44	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
Capy-team	3	0.20	0.29	0.16	0.10	0.10	0.14	0.20	0.26	0.14	0.10	0.12	0.10
DS-GT	1	0.90	0.92	0.52	0.00	0.00	0.12	0.20	0.18	0.20	0.10	0.12	0.17

should be made in this line in future work. The limitation of our system in the decision-based evaluation lies in the decision threshold τ, which could not be fine-tuned using contextual data during the training stage.

6. Conclusions

In this article we deal with early detection of user risk level of depression in social media threads. Even though previous approaches focused on posts written by the target user, in this task we count on the posts written in context by the target user and also by the interacting users. The scenario emulates a real continuous monitoring environment with full conversational context. The research question rests on the means to leverage contextual information, provided by other posts, together with target user posts in what it comes to seize the risk on the target user. The system has to process the thread in chronological subsequences of messages and has to provide two outcomes: the high risk alarm (or no alarm) decision together with the confidence on the decision made. The assessment comprises decision accuracy and earliness together with confidence ranking. An added challenge in this tasks rests on the fact that the training data consisted on mere single-user posts without context.

We contribute with an early depression detection approach that integrates individual user risk signals with conversational context. The results from the ranking-based evaluation demonstrate the validity of our approach, as we were able to rank users by their risk level in a highly competitive manner, consistently placing among the top participants. However, in the decision-based evaluation our precision and F_1 scores were impacted by the choice of a relatively low decision threshold. This conservative setting, determined without contextual data during training, increased the number of false positives and reduced performance on classification metrics.

Future work will explore adaptive thresholding techniques that leverage contextual information to optimally balance precision and recall, as well as more advanced context-encoding mechanisms to further enhance early depression detection.

Acknowledgments

This work was partially funded by the Spanish Ministry of Science and Innovation (EDHIA PID2022-

136522OB-C22) and by the Basque Government (IXA IT-1570-22). Besides, this work was elaborated within the framework of LOTU (TED2021-130398B-C22) funded by MCIN/AEI/10.13039/501100011033, European Comission (FEDER), and by the European Union "NextGenerationEU"/PRTR. The first author is a recipient of a grant from the Spanish Ministry of Education's *Formación de Profesorado Universitario* program (FPU23/01068).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of risk 2022: Early risk prediction on the internet, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 233–256.

[2] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 294–315.

[3] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 73–92.

[4] H. Fabregat, D. Deniz, A. Duque, L. Araujo, J. Martínez-Romo, NLP-UNED at eRisk 2024: Approximate Nearest Neighbors with Encoding Refinement for Early Detecting Signs of Anorexia, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024, pp. 813–824.

[5] O. Riewe-Perła, A. Filipowska, Combining Recommender Systems and Language Models in Early Detection of Signs of Anorexia, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, Grenoble, France, 2024.

[6] M. Kula, Metadata embeddings for user and item cold-start recommendations, in: T. Bogers, M. Koolen (Eds.), Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015., volume 1448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 14–21. URL: http://ceur-ws.org/Vol-1448/paper4.pdf.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[8] A. M. Mármol Romero, A. Moreno Muñoz, F. M. Plaza del Arco, M. D. Molina González, M. T. Martín Valdivia, L. A. Ureña López, A. Montejo Ráez, Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish, Procesamiento del lenguaje natural 71 (2023) 329–350.

[9] A. M. Mármol Romero, A. Moreno Muñoz, F. M. Plaza del Arco, M. D. Molina González, M. T. Martín Valdivia, L. A. Ureña López, A. Montejo Ráez, Overview of mentalriskes at iberlef 2024: Early detection of mental disorders risk in spanish, Procesamiento del lenguaje natural 73 (2024) 435–448.

[10] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation

Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

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

[12] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2016, pp. 28–39.

[13] X. Larrayoz, N. Lebena, A. Casillas, A. Pérez, Representation Exploration and Deep Learning Applied to the Early Detection of Pathological Gambling Risks, in: CLEF (Working Notes), 2023, pp. 693–705.

[14] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, Uned-nlp at erisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:251471984.

[15] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, arXiv preprint arXiv:2402.05672 (2024).