Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 425-440 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of MentalRiskES at IberLEF 2025: Early Detection of Addiction Risk in Spanish

Resumen de la Tarea MentalRiskES en IberLEF 2025: Detección Precoz de Adicciones en Español

Alba María Mármol-Romero,[1] Pablo Álvarez-Ojeda,[1] Adrián Moreno-Muñoz,[1] Flor Miriam Plaza-del-Arco,[2] M. Dolores Molina-González,[1] M. Teresa Martín-Valdivia,[1] L. Alfonso Ureña-López,[1] Arturo Montejo-Ráez[1]
[1]University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
[1]{amarmol, ammunoz, mdmolina, maite, laurena, amontejo}@ujaen.es,
[2]{f.m.plaza.del.arco@liacs.leidenuniv.nl}

Abstract: This paper presents the MentalRiskES shared task organized at IberLEF 2025, as part of the 41st International Conference of the Spanish Society for Natural Language Processing. This task aims to promote the early detection of mental risk disorders in Spanish. We propose two detection tasks: Task 1 for risk detection of gambling disorders, Task 2 for risk detection of gambling disorders but determining the type of addiction. Furthermore, we asked participants to submit measurements of carbon emissions for their systems, emphasizing the need for sustainable natural language processing practices. In this third edition, 27 teams registered, 13 submitted results, and 12 presented papers. Most teams experimented with Transformers, including features, data augmentation, and preprocessing techniques.

Keywords: mental disorder risk detection, early detection of anxiety, early detection of depression, early detection of eating disorders.

Resumen: Este artículo presenta la tarea MentalRiskES en IberLEF 2025, como parte de la 41º edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. El objetivo de esta competición es promover la detección temprana de trastornos mentales en español. Proponemos dos tareas de detección precoz: Tarea 1 para detección de riesgo de trastornos por juego, Tarea 2 para detección de riesgo de trastornos por juego pero determinando el tipo de juego relacionado. Además, pedimos a los participantes que enviaran mediciones de las emisiones de carbono de sus sistemas, haciendo hincapié en la necesidad de prácticas sostenibles de procesamiento del lenguaje natural. En esta tercera edición, 27 equipos se registraron, 13 enviaron predicciones y 12 presentaron artículos. La mayoría experimentó con Transformers, incluyendo características, ampliando datos y técnicas de preprocesamiento.

Palabras clave: detección precoz de trastornos mentales, detección precoz de adición al juego, detección precoz de ludopatía, detección precoz de adición a videojuegos.

1 Introduction

The digital age has vastly increased engagement with online platforms. Studies show that heavy social media use by adolescents is significantly associated with higher levels of depression, anxiety, and other mental health problems (Agyapong-Opoku, Agyapong-Opoku, and Greenshaw, 2025). At the same time, Internet access has made online gaming and gambling more ubiquitous. The World Health Organization (WHO) notes that most people who game or gamble experience no serious issues, but a small minority develop addictive disorders with substantial impairment [1]. In particular, repetitive gambling behavior can escalate into a formal gambling disorder characterized

[1] https://www.who.int/health-topics/addictive-behaviour

ISSN 1135-5948 DOI 10.26342/2025-75-31

by significant distress and functional impairment.

In Spain, gambling—both online and offline—is a widespread practice. According to the *Informe sobre Adicciones Comportamentales 2024* (Observatorio Español de las Drogas y las Adicciones (OEDA), 2024), 53.8% of the population aged 15 to 64 reported having gambled in the last 12 months, with 5.5% engaging specifically in online gambling. Notably, gambling behavior begins at an early age: the average age of first-time gambling is 14.7 years for online gambling and 14.8 years for offline gambling, highlighting the particular vulnerability of adolescents. Among students aged 14 to 18 who reported gambling online in 2023, over half (51.5%) did so through gambling-related video games. Furthermore, the use of cryptocurrencies and trading emerged as a new gambling modality: 26.8% of online-gambling students reported participating in gambling activities involving cryptocurrencies or trading platforms.

In recent years, Natural Language Processing (NLP) and deep learning have been increasingly applied to identify early signs of mental health problems from user-generated texts. Notably, the CLEF eRisk Lab has organized tasks for early detection of depression, self-harm, and early detection of pathological gambling (Parapar et al., 2021; Parapar et al., 2022; Parapar et al., 2023). However, most of these initiatives have focused on English data, leaving languages like Spanish underexplored.

This paper describes the third edition of a novel task on early risk identification of mental disorders in Spanish comments from social media sources organized within the Iberian Languages Evaluation Forum (González-Barba, Chiruzzo, and Jiménez-Zafra, 2025). The first edition (Mármol-Romero et al., 2023) took place two years ago in the Iberian Languages Evaluation Forum (IberLEF) as part of the International Conference of the Spanish Society for Natural Language Processing (SEPLN) 2023. The second edition (Mármol-Romero et al., 2024b) took place last year in the same congress and forum, SEPLN 2024. The task was resolved as an online problem, meaning that participants had to detect a potential risk as early as possible in a continuous stream of data. Therefore, the performance not only depends on the accuracy of the systems but also on how fast the problem is detected. These dynamics are reflected in the design of the tasks and the metrics used to evaluate participants. For this third edition, we propose two novel tasks. The first subtask is about the detection of gambling disorders, and the second subtask consists of detecting the scope where the addiction takes place.

2 Tasks

In this section, we describe the different tasks proposed in the third edition of the competition.

2.1 Task 1: Risk Detection of Gambling Disorders

Detect if a user is at high risk or low risk of developing a gambling-related disorder based on their messages. This is a binary classification task with two possible labels: high risk (label = 1) or low risk (label = 0).

2.2 Task 2. Type of Addiction Detection

The first part of this task consists of detecting if a user is at high risk or low risk of developing a gambling-related disorder. This is the same as the task 1 described before. In addition to detecting the risk, participants must detect the specific type of addiction associated with the disorder, interpreting the *type* as the scope or topic where the addiction may emerge. Available types are:

- **Betting**: a gambling activity where individuals place bets on sports-related events, aiming to win money based on the outcome.
- **Online Gaming**: gambling activity where individuals participate in traditional games of chance, such as roulette, blackjack, or slot machines, wagering money either in person or online.
- **Trading and crypto**: gambling activity where individuals engage in speculative investments, particularly in cryptocurrencies, facing uncertainty regarding financial gains or losses.
- **Lootboxes**: a gambling activity where individuals purchase virtual items in video games that contain randomized content, often using real-world money, introducing an element of financial risk.

2.3 Evaluation measures

Tasks are evaluated based on their specific definitions. We assess a system's performance using two primary criteria: **absolute classification** and **early detection effectiveness**. Table 3 in Appendix A shows the evaluation perspective for each task (each task needs a different way to be evaluated due to the nature of the decisions requested) and the metrics used to evaluate them.

2.3.1 Classification-based evaluation

This evaluation method focuses on binary or multi-class classification decisions made by the participating systems for each user. The objective is to determine whether a user is at high risk of experiencing a gambling disorder. To evaluate Tasks 1 and 2, we used classical metrics such as accuracy, macro-precision, macro-recall, and macro-F1. These metrics assess the systems' final predictions after analyzing all posts from each subject in the dataset. For system ranking purposes, we selected the **macro-F1** metric as the primary criterion.

2.3.2 Latency-based evaluation

We draw on the established framework from eRisk (Parapar et al., 2021) to derive metrics for measuring the early detection of positive subjects by the participating systems. For evaluating Tasks 1 and 2, we used early risk evaluation metrics such as ERDE (Losada and Crestani, 2016) (ERDE5 and ERDE30), latencyTP, speed, and latency-weighted F1 (Sadeque, Xu, and Bethard, 2018). Given the short length of messages in our dataset, we determined that a larger number of messages is necessary for accurate early detection, leading us to prioritize the **ERDE30** metric for system ranking.

2.3.3 Efficiency metrics

Efficiency metrics are intended to measure the impact of the system in terms of resources needed and environmental issues. These metrics are not used to rank the system but to recognize their efficiency and, accordingly, if the carbon footprint is environmentally friendly. So, we use metrics to measure the level of carbon emissions produced by a system while it is predicting. We aim to recognize systems capable of performing tasks with minimal resource demand. This allows us to identify technologies that can operate on mobile devices or personal computers and those with the lowest energy consumption and carbon footprint. To achieve this, each final prediction will include the following information:

- Minimum, maximum, mean, and variance of prediction time.
- Minimum, maximum, mean, and variance of CO2 emissions generated per prediction.
- Minimum, maximum, mean, and variance of energy consumption per CPU/GPU (kW) during prediction.
- Minimum, maximum, mean, and variance of RAM energy consumption (kW) during prediction.
- Minimum, maximum, mean, and variance of total energy consumption (kW) combining CPU, GPU, and RAM.
- Number and models of CPUs/GPUs used, the total RAM size required and the 3-letter alphabet ISO Code of the respective country.

Participants used the CodeCarbon package[2] to track emissions, measured in kilograms of CO2-equivalents (CO2eq), to estimate the carbon footprint of their system predictions.

3 Dataset

In this edition, for tasks 1 and 2 of the MentalRiskES shared task, we utilized threads of messages extracted from the PRECOM-SM corpus (Álvarez-Ojeda et al., 2025). This corpus was specifically compiled to study gambling in Spanish social media and includes text from various online platforms such as Telegram, Twitch, Reddit, and Ludopatia.org. The messages in this dataset follow a similar structure to the dataset used in previous years' MentalRiskES shared tasks (Mármol-Romero et al., 2024a).

For this year's tasks, the dataset was specifically labeled to indicate either a low or high risk of suffering from a gambling disorder. This labeling was based on a primary hypothesis that users with a higher number of messages are more likely to be at a higher risk of a disorder, reflecting more frequent engagement, while users with fewer messages

[2] https://mlco2.github.io/codecarbon/index.html

are assumed to be at lower risk. This hypothesis aligns with existing research indicating a correlation between active participation in online communities and at-risk or problem gambling behavior (Kuss and Griffiths, 2012; Jayemanne et al., 2021). To ensure a representative set of users for the competition, we worked to select users that represent both high and low interaction frequencies.

3.1 Data Selection and Splitting

To construct the training, trial, and test sets for each addiction type (betting, online gaming, trading/crypto, lootboxes), a specific subject and message selection strategy was employed. This strategy aimed to balance the dataset while preventing participants from inferring risk levels solely based on message count. The process involved:

1. Random Subject Selection: For each addiction type, N subjects were randomly selected from the "low risk" group, and an equal number (N) from the "high risk" group. This ensured a balanced distribution of risk categories within the dataset.

2. Message count adjustment for high-risk subjects: To prevent message volume from serving as an unintended proxy for risk level, the number of messages retained for each high-risk subject was determined by sampling from a normal distribution. The mean and standard deviation of this distribution were calculated based on the message counts of the low-risk subjects. This step introduced controlled variability while maintaining plausible message counts across both groups.

3. Selection of most recent messages: Once the target message count X was determined for a high-risk subject, the final dataset retained their X most recent messages. This choice prioritized the users' latest linguistic behavior, which is particularly relevant for identifying current signs of problematic use.

This selection strategy was designed to yield a realistic and challenging dataset, encouraging participants to rely on linguistic cues rather than superficial metadata to assess addiction risk.

3.2 Dataset statistics

A single dataset is presented, consisting of user messages coming from different platforms where people chat on topics related to gambling behaviour. Each user is labeled with a risk level (high or low) and a type of addiction (e.g., betting, online gaming, trading/crypto, lootboxes). Each user and their corresponding messages were split into three sets: (1) trial – used for system validation, (2) train – used for training models, and (3) test – used for evaluation. Table 1 summarizes the distribution of subjects and messages across each addiction type, risk level, and data partition.

The train and trial sets were sent to the participant as a .zip file containing JSON files. Each JSON contained a history of messages for a subject with the attributes: (1) id message, to identify the message; (2) message, the text message; (3) date, the date and time when the message was sent to the group; and (4) platform, the platform from which the data was extracted. On the other hand, to test the server, the trial set, again, and the test set was sent by a get request to a server whose response was a JSON file that contained a collection of messages from a lot of different subjects in one specific round. This process is repeated until all the messages from all the subjects are sent. The attributes for each JSON were: (1) id message, to identify the message; (2) nick, to identify the subject; (3) round, to identify the round; (4) message, the text message; (5) platform, the platform from which the data was extracted and (6) date, the date and time when the message was sent to the group.

4 Baselines

To establish a baseline benchmark for the tasks, we performed experiments using two different Transformer-based models. We experimented with Spanish pre-trained models such as RoBERTa Base from the MarIA project (Fandiño et al., 2022), and RoBERTuito, a pre-trained language model for user-generated content in Spanish (Pérez et al., 2021). These models have demonstrated favourable results in Spanish tasks. In addition, RoBERTa Base,[3] and RoBERTuito[4] are available at the HuggingFace models' hub.[5]

[3] `PlanTL-GOB-ES/RoBERTa-base-bne`

[4] `pysentimiento/robertuito-base-cased`

[5] `https://huggingface.co`

		Trial		Train		Test	
		Subj	Msg	Subj	Msg	Subj	Msg
Betting	High risk	1	125	45	5,074	15	1,759
	Low risk	1	135	40	4,492	20	2,237
Online Gaming	High risk	1	24	54	1,406	20	520
	Low risk	1	1	50	1,297	24	640
Trading and crypto	High risk	1	80	60	4,465	35	2,546
	Low risk	1	71	75	5,542	20	1,486
Lootboxes	High risk	0	0	13	107	13	111
	Low risk	1	8	13	108	13	108

Table 1: Distribution of subjects (Subj) and messages (Msg) by type of addiction related to gambling disorder and risk level across partitions.

For experiments in tasks 1 and 2, we trained using the training set, used the trial set for early stopping, and evaluated using the test set. The experiments with Transformer used default hyper-parameters, however, we applied a fine-tuning that is specified in Table 2 and added a TrainerCallback to handle early stopping. All the training and evaluation experiments were performed on a node equipped with 2 NVIDIA RTX 4000 SFF Ada Generation.

Hyperparameters	Value
Learning Rate	5e-5
Weight Decay	0
Batch size	8
Seed	42
Max length	512
Number of train epochs	20
Early stopping patience	10

Table 2: Baselines training details for transformers-based experiments.

4.1 Task 1: Risk Detection of Gambling Disorders

In the HuggingFace transformer training arguments, the number of labels was set to 2, and the problem type was set to single-label classification. Early stopping was set to stop when the highest value in the macro-averaged F1 score was reached. It needed 6 epochs for RoBERTuito and 7 for RoBERTa Base.

4.2 Task 2: Type of Addiction Detection

For the first level, we use the same predictions calculated for task 1. For the second level, in the HuggingFace transformer training arguments, the number of labels was set to 4, and the problem type was set to single-label classification. Early stopping was set to stop when the highest value in the macro-averaged F1 score was reached. It needed 2 epochs for both models, RoBERTuito and RoBERTa Base.

5 *Participant approaches*

- **wangkongqiang** (Wang, 2025). The team participated in Task 1, their approach involved concatenating each user's messages and augmenting the training data by splitting these into halves and thirds to simulate early detection. They compared two main strategies: fine-tuning pre-trained RoBERTa models and training regression algorithms on sentence embeddings.
- **VerbaNexAI Lab** (Jimenez et al., 2025). This team participated in both tasks. Their approach combined transformer-based Spanish BERT embeddings with traditional machine learning classifiers, using a pipeline of text preprocessing, class balancing, and systematic model selection.
- **SoloResearch** (Sologuestoa et al., 2025). This team participated in both tasks. Their approach centered on lightweight, resource-efficient models, employing Google's multilingual embeddings and data augmentation via generative LLMs to address limited annotated data. For early risk detection, they used a Bi-LSTM with dual (learned and lexicon-based) attention and GroupDRO loss, while addiction type classification was tackled with a hierarchical Bi-LSTM. Dynamic thresholding and segment-specific data augmentation were key methodological components, with all systems designed

to run without GPU acceleration.

- **purple_john** (Damov, Ioan, and Schlupek, 2025). The team participated in Task 1, exploring a wide range of models and transformer-based architectures using both Spanish and English pretrained embeddings. Their pipeline involved standard text preprocessing and did not use external or augmented data.
- **HULAT_UC3M** (Campos-Molina and Martínez, 2025). This team participated in Task 1 with three machine learning-based approaches. Two are based on SVMs with message embeddings, while the third combined data augmentation, sentence vectorization, and a Random Forest model, achieving the best result among the three strategies.
- **PUXai** (Phuc and Thin, 2025). This team participated in Tasks 1 and 2 using pretrained Spanish language models with data augmentation and optimization techniques. For Task 1, they used RoBERTuito with an LSTM and GroupDRO; for Task 2, roberta-base-bne with back-translation and Optuna tuning.
- **PLN_PPM_ISB** (Molina and Bedmar, 2025). This team participated in both tasks using a mix of classic and transformer-based models. For Task 1, they applied SVMs, including one enhanced with zero-shot scores from a RoBERTa-based model, as well as a BERT-based Spanish model with data augmentation. For Task 2, they applied data augmentation to both SVM and transformer models to address class imbalance. Their approach combined traditional ML, zero-shot inference, and pretrained language models to adapt to the informal and varied nature of user-generated content.
- **UNSL** (Thompson and Errecalde, 2025). This team only participated in Task 1. They proposed three methods grounded in a CPI+DMC framework, treating predictive effectiveness and decision-making speed as separate objectives. The models used were SS3, BERT with an extended vocabulary, and SBERT, complemented by decision policies based on historical user behaviour.
- **UC3Mental** (Rodriguez, Zubasti, and Saiz, 2025). This team participated in both tasks, applying three approaches: an SVM baseline, a two-stage Random Forest plus SVM pipeline leveraging addiction type for risk detection, and a BERT-based transformer model, all using concatenated and preprocessed user messages.
- **ELiRF-UPV** (Casamayor et al., 2025). This team participated in both tasks. They implemented three approaches for both tasks, one is based on SVM classifier and the other two are based on transforms architectures (RoBERTa and Longformer). In addition, for Task 1, the authors extended the training data by using an augmentation data technique.
- **MCDI** (Herrera and Tellez, 2025). This team develops two approaches for tackling Task 2 of MentalRiskES 2025. One is based on bag-of-words model and the other on RoBERTuito-based model.
- **I2C-UHU-Rigel** (Moreno, Vázquez, and Álvarez, 2025). This team participated in both tasks. They implemented three approaches for both tasks based on fine-tuning pre-trained transformer architectures (BERT, XML-RoBERTa and distilBERT). In addition, for Task 2, the authors extended the training data by using an augmentation data technique with back translation, and they experimented with chunk-based segmentation to effectively handle message history.

6 Results

As mentioned in Section 2.3, tasks are evaluated according to how the task is defined. We evaluate a system according to its performance in terms of **absolute classification** and in terms of **early detection effectiveness** for classification tasks. Section 2.3 provides an overview of the evaluation metrics used for each task.

6.1 Task 1

Participant results and the baselines proposed are shown in Table 4 (absolute classification) and Table 5 (early detection).

In terms of absolute classification, the top-performing team, UNSL (Run 2), achieved the highest Macro-F1 score of 0.567, outperforming the baseline models. According to the ranking, UNSL (Run 0) and I2C-UHU-Rigel (Run 1) followed, both also surpassing the baselines, including Roberta Base and Robertuito. Notably, the highest Macro-Precision (0.657) was obtained by the baseline model Robertuito, while the highest Macro-Recall (0.574) was achieved by UNSL (Run 0). In terms of early detection, the top-performing team, PLN_PPM_ISB, achieved the lowest ERDE30 value of 0.242, surpassing the baseline models. Following in the ranking, UC3Mental and VerbaNexAI Lab also obtained competitive ERDE30 scores. Notably, VerbaNexAI Lab achieved the best ERDE5 value (0.274), indicating strong performance in very early detection. Among the baselines, Robertuito stands out with the highest latency-weighted F1 score (0.685).

6.2 Task 2

Participant results and the baselines proposed are shown in Table 6 (absolute classification), Table 7 (early detection) and Table 8 (absolute classification for type).

In terms of absolute classification for Task 2, the top-performing team, MCDI, achieved the highest Macro-F1 score of 0.589 across all runs, surpassing the baseline models. HULAT_UC3M and ELiRF-UPV also obtained competitive results, while the baseline Robertuito achieved the highest Macro-Precision (0.657). For early detection, the best ERDE30 value (0.242) was obtained by PLN_PPM_ISB, outperforming the baselines. Notably, VerbaNexAI Lab achieved the best ERDE5 score (0.274), indicating strong performance in very early detection, and Robertuito stood out among the baselines with the highest latency-weighted F1 score (0.685). In the multi-class classification scenario, the highest Macro-F1 (0.927) was achieved by PLN_PPM_ISB, followed by ELiRF-UPV and SoloResearch, while both Robertuito and Roberta Base maintained solid baseline performance.

7 *Discussion*

Moderate but consistent performance across teams suggests that current approaches could support human experts rather than replace clinical assessment. Integration into existing mental health support systems could provide valuable early warning capabilities, particularly when combined with other risk indicators.

The overwhelming adoption of transformer-based architectures across participating teams confirms their effectiveness for Spanish mental health risk detection. However, the variation in performance among transformer-based approaches suggests that architecture choice alone is insufficient. Successful teams typically combine transformers with domain-specific enhancements such as:

- Data augmentation techniques: Several teams used generative LLMs for data expansion and back-translation methods.
- Specialized attention mechanisms: Bi-LSTM with dual attention (SoloResearch) showed competitive performance.
- Hierarchical approaches: Multi-level classification strategies proved effective for complex categorization tasks.

In the following, a detailed analysis is discussed for each task and subtask.

7.1 Classification-based evaluation in Task 1.

The results show moderate but promising performance across participating teams. The top-performing team, UNSL (Run 2), achieved a Macro-F1 score of 0.567, which represents a substantial improvement over the baseline models (RoBERTuito: 0.428, RoBERTa Base: 0.342). This performance gap suggests that task-specific adaptations and methodological innovations can significantly enhance detection capabilities beyond general pre-trained models.

Notably, the performance distribution shows a clear distinction between specialized approaches and baseline methods. Teams that implemented sophisticated techniques such as data augmentation (wangkongqiang), hierarchical classification (SoloResearch), and ensemble methods (multiple teams) generally outperformed simpler approaches. However, the moderate absolute scores (ranging from 0.342 to 0.567 for Macro-F1) indicate that gambling disorder detection remains a challenging task, likely due to the subtle linguistic patterns and the potential similarity between high-risk and low-risk language use.

7.2 Latency-based evaluation in Task 1

The early detection metrics reveal critical insights about the temporal aspects of risk identification. PLN_PPM_ISB achieved the best ERDE30 score of 0.242, while VerbaNexAI Lab excelled in very early detection with an ERDE5 score of 0.274. These results indicate that effective early detection is achievable, though the performance varies significantly depending on the amount of available data.

The latency-weighted F1 scores (ranging from 0.377 to 0.685) suggest that there is often a trade-off between detection speed and accuracy. Teams that achieved better early detection metrics sometimes showed lower absolute classification performance, highlighting the inherent tension between timeliness and precision in early warning systems.

7.3 Classification-based evaluation in Task 2

For binary classification (risk detection), MCDI achieved the highest performance with a Macro-F1 of 0.589, slightly outperforming Task 1 results. This improvement might be attributed to the additional training signal provided by the multi-class annotation, which could help models better understand the underlying patterns.

The multi-class classification results (Table 8) show remarkably higher performance, with PLN_PPM_ISB achieving a Macro-F1 of 0.927. This substantial difference suggests that distinguishing between different types of gambling-related content (betting, online gaming, trading/crypto, lootboxes) is considerably easier than detecting risk levels. This finding has important implications for system design, indicating that content categorization could serve as an intermediate step to improve risk assessment.

7.4 Latency-based evaluation in Task 2

The early detection metrics reveal critical insights about temporal aspects of risk identification. PLN_PPM_ISB achieved the best ERDE30 score of 0.242, while VerbaNexAI Lab excelled in very early detection with an ERDE5 score of 0.274. These results indicate that effective early detection is achievable, though the performance varies significantly depending on the amount of available data.

The latency-weighted F1 scores (ranging from 0.377 to 0.685) suggest that there is often a trade-off between detection speed and accuracy. Teams that achieved better early detection metrics sometimes showed lower absolute classification performance, highlighting the inherent tension between timeliness and precision in early warning systems.

7.5 Efficiency analysis

The success of resource-efficient approaches also suggests potential for deployment in varied computational environments, from research institutions to community organizations with limited technical infrastructure.

8 Conclusions

The MentalRiskES 2025 shared task results reveal important insights about the challenges and opportunities in early detection of gambling disorders in Spanish social media text. This third edition attracted significant participation with 27 registered teams, 13 successful submissions, and 12 presented papers, demonstrating the growing interest in mental health risk detection for Spanish-language content.

The results demonstrate both the potential and challenges of automated gambling disorder risk detection in Spanish social media. While no system achieved performance levels suitable for autonomous clinical decision-making, the consistent improvements over baseline approaches and the successful implementation of early detection capabilities suggest meaningful progress toward practical applications. The integration of sustainability considerations and the diversity of methodological approaches reflect a maturing field that balances technical innovation with practical and ethical considerations.

As future work, we plan to explore new editions of MentalRiskES focused not only on risk detection but also on the identification of specific symptoms and potential treatment strategies. This shift will involve the use of large language models (LLMs) to support more nuanced, personalized, and clinically relevant analyses of mental health content, moving beyond detection to intervention support.

Acknowledgments

Our sincere thanks to the organizers of the IberLEF workshop, who, once again this year, have allowed us to organize the shared task.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

Agyapong-Opoku, N., F. Agyapong-Opoku, and A. J. Greenshaw. 2025. Effects of social media use on youth and adolescent mental health: A scoping review of reviews. *Behavioral Sciences*, 15(5):574.

Álvarez-Ojeda, P., M. V. Cantero-Romero, A. Semikozova, and A. Montejo-Ráez. 2025. The precom-sm corpus: Gambling in spanish social media. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 17–28.

Campos-Molina, J. and P. Martínez. 2025. Hulat-uc3m at task1@mentalriskes 2025: Detecting gambling disorders using machine learning approaches. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Casamayor, A., V. Ahuir, A. Molina, and L.-F. Hurtado. 2025. ELiRF-UPV at MentalRiskES 2025: Spanish Longformer for Early Detection of Gambling Addiction Risk. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Damov, C., A. Ioan, and C. Schlupek. 2025. Early Detection of Gambling Addiction Risk in Spanish: purple_john at MentalRIskES2025. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org.*

Herrera, A. and E. S. Tellez. 2025. INFOTEC-MCDI at MentalRiskES: Gambling Risk Early Prediction in Social Media Users Through Bag of Word and BERT Ensembles. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Jayemanne, D., S. Chillas, J. Moir, A. Rocha, F. Simpson, and H. Wardle. 2021. Loot boxes and digital gaming: a rapid evidence assessment.

Jimenez, J. D., J. E. Serrano, J. C. Martinez-Santos, and E. Puertas. 2025. VerbaNexAI at MentalRiskES 2025: Early Detection of Gambling Disorders using Transformer Architectures and Machine Learning Models. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Kuss, D. J. and M. D. Griffiths. 2012. Internet gaming addiction: A systematic review of empirical research. *International journal of mental health and addiction*, 10:278–296.

Losada, D. and F. Crestani. 2016. A test collection for research on depression and language use. volume 9822, pages 28–39, 09.

Mármol-Romero, A. M., A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Ráez. 2024a. MentalRiskES: A new corpus for early detection of mental disorders in Spanish. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11204–11214, Torino, Italia, May. ELRA and ICCL.

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

Mármol-Romero, A. M., A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Raéz. 2023. Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish. *Procesamiento del Lenguaje Natural*, 71:329–350.

Molina, P. P. and I. S. Bedmar. 2025. PLN_PPM_ISB at MentalRiskES 2024: Detection of Gambling Disorders and Type of Addiction. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Moreno, A. L. G., J. M. Vázquez, and V. P. Álvarez. 2025. I2C-UHU-Rigel at MentalRiskES 2025: Detection of Gambling Disorder Risk in Spanish using Transformer-Based Models. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Observatorio Español de las Drogas y las Adicciones (OEDA). 2024. Informe sobre adicciones comportamentales y otros trastornos adictivos 2024. Ministerio de Sanidad, Delegación del Gobierno para el Plan Nacional sobre Drogas.

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

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2022. Overview of erisk at clef 2022: Early risk prediction on the internet (extended overview). In *CEUR Workshop Proceedings (CEUR-WS. org)*.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *CoRR*, abs/2111.09453.

Phuc, N. X. and D. V. Thin. 2025. Puxai at mentalriskes 2025: Robertuito with bidirectional lstm for early gambling disorder detection. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Rodriguez, M., P. Zubasti, and M. Saiz. 2025. UC3Mental at MentalRiskES 2025: RF-SVM Ensemble Approach for Early Detection of Mental Health Risks Using NLP. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Sadeque, F., D. Xu, and S. Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Sologuestoa, I., X. Larrayoz, M. Oronoz, and A. Pérez. 2025. Data Augmentation via Generative LLMs for the Detection of Gambling Disorders and Type of Addiction in Social Media Threads. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Thompson, H. and M. Errecalde. 2025. Tackling a Challenging Corpus for Early Detection of Gambling Disorder: UNSL at MentalRiskES 2025. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Wang, K. 2025. wangkongqiang at MentalRiskES@IberLEF 2025: Early Detection of Mental Disorders Risk in Spanish. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

A Metrics used in evaluating submissions

Table 3 shows the evaluation perspective for each task (each task needs a different way to be evaluated due to the nature of the decisions requested) and the metrics used to evaluate them.

B Participant Results

Tasks	Evaluation perspective	Metrics
1, 2	Absolute multi-class and binary classification	Accuracy, Macro-P, Macro-R **Macro-F1**
2	Absolute multi-class classification	Accuracy, Macro-P, Macro-R **Macro-F1**, Micro-P Micro-R Micro-F1
1, 2	Early detection in multi-class and binary classification	ERDE5, **ERDE30**, latencyTP, speed, latency-weightedF1

Table 3: Metrics used in evaluating submissions to MentalRiskES 2025 tasks. The reference metric (for submission ranking) for that evaluation is in bold.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	UNSL	2	0.569	0.568	0.567	**0.567**
2	UNSL	0	**0.581**	0.586	**0.574**	0.563
3	I2C-UHU-Rigel	1	0.556	0.555	0.553	0.551
4	ELiRF-UPV	1	0.550	0.564	0.556	0.540
5	ELiRF-UPV	2	0.538	0.543	0.542	0.534
6	ELiRF-UPV	0	0.538	0.535	0.535	0.533
7	PLN_PPM_ISB	2	0.538	0.535	0.534	0.532
8	UC3Mental	1	0.513	0.507	0.506	0.495
9	HULAT_UC3M	2	0.488	0.488	0.488	0.488
10	PLN_PPM_ISB	1	0.569	0.636	0.554	0.484
11	SoloResearch	2	0.506	0.497	0.498	0.475
12	NLP-UNED	0	0.531	0.533	0.519	0.468
13	SoloResearch	1	0.500	0.486	0.490	0.455
14	NLP-UNED	2	0.494	0.476	0.483	0.446
15	SoloResearch	0	0.494	0.476	0.483	0.446
16	UNSL	1	0.475	0.459	0.467	0.444
17	PLN_PPM_ISB	0	0.550	0.632	0.534	0.436
18	UC3Mental	2	0.550	0.632	0.534	0.436
19	NLP-UNED	1	0.506	0.487	0.493	0.429
20	Robertuito		0.550	**0.657**	0.533	0.428
21	PUXai	0	0.538	0.622	0.520	0.403
22	PUXai	2	0.525	0.539	0.508	0.396
23	UC3Mental	0	0.506	0.465	0.490	0.385
24	I2C-UHU-Rigel	2	0.481	0.490	0.498	0.371
25	PUXai	1	0.525	0.594	0.507	0.367
26	purple_john	0	0.519	0.509	0.500	0.353
27	VerbaNexAI Lab	0	0.519	0.259	0.500	0.342
28	VerbaNexAI Lab	1	0.519	0.259	0.500	0.342
29	VerbaNexAI Lab	2	0.519	0.259	0.500	0.342
30	I2C-UHU-Rigel	0	0.519	0.259	0.500	0.342
31	Roberta Base		0.519	0.259	0.500	0.342
32	wangkongqiang	2	0.519	0.259	0.500	0.342
33	wangkongqiang	1	0.519	0.259	0.500	0.342
34	wangkongqiang	0	0.519	0.259	0.500	0.342
35	purple_john	1	0.519	0.259	0.500	0.342
36	HULAT_UC3M	0	0.513	0.258	0.494	0.339
37	purple_john	2	0.513	0.258	0.494	0.339
38	HULAT_UC3M	1	No valid run			

Table 4: Classification-based evaluation in Task 1. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	PLN_PPM_ISB	0	0.316	**0.242**	**2**	**0.990**	0.683
2	PLN_PPM_ISB	1	0.412	0.248	**3**	0.981	0.680
3	UC3Mental	2	0.415	0.249	3	0.981	0.676
4	VerbaNexAI Lab	0	**0.274**	0.250	**2**	**0.990**	0.677
5	VerbaNexAI Lab	1	**0.274**	0.250	**2**	**0.990**	0.677
6	VerbaNexAI Lab	2	**0.274**	0.250	**2**	**0.990**	0.677
7	I2C-UHU-Rigel	0	0.301	0.250	**2**	**0.990**	0.677
8	wangkongqiang	2	0.354	0.250	3	0.981	0.670
9	wangkongqiang	1	0.333	0.250	3	0.981	0.670
10	wangkongqiang	0	0.332	0.250	3	0.981	0.670
11	purple_john	1	0.371	0.250	3	0.981	0.670
12	Robertuito		0.329	0.252	**2**	**0.990**	**0.685**
13	purple_john	0	0.377	0.253	3	0.981	0.667
14	Roberta Base		0.280	0.256	**2**	**0.990**	0.676
15	HULAT_UC3M	0	0.363	0.256	2	**0.990**	0.671
16	purple_john	2	0.382	0.256	3	0.981	0.665
17	PUXai	1	0.577	0.262	5	0.962	0.657
18	PUXai	2	0.377	0.283	3	0.981	0.662
19	UC3Mental	0	0.445	0.283	3	0.981	0.645
20	UNSL	0	0.515	0.284	5	0.962	0.628
21	PUXai	0	0.434	0.302	3	0.981	0.673
22	UC3Mental	1	0.575	0.334	7	0.942	0.556
23	NLP-UNED	0	0.587	0.336	11	0.909	0.592
24	PLN_PPM_ISB	2	0.454	0.340	4	0.976	0.570
25	NLP-UNED	1	0.595	0.352	10	0.914	0.584
26	NLP-UNED	2	0.617	0.374	10	0.914	0.556
27	UNSL	2	0.639	0.389	17	0.848	0.506
28	ELiRF-UPV	2	0.600	0.394	14	0.876	0.432
29	ELiRF-UPV	1	0.579	0.402	14	0.876	0.412
30	ELiRF-UPV	0	0.649	0.410	21	0.810	0.470
31	SoloResearch	0	0.639	0.424	11	0.904	0.550
32	I2C-UHU-Rigel	1	0.600	0.432	19	0.829	0.496
33	SoloResearch	1	0.674	0.458	20	0.820	0.501
34	HULAT_UC3M	2	0.624	0.472	20	0.820	0.400
35	UNSL	1	0.707	0.476	21	0.810	0.467
36	SoloResearch	2	0.691	0.487	26	0.769	0.464
37	I2C-UHU-Rigel	2	0.533	0.516	12	0.895	0.096
38	HULAT_UC3M	1	No valid run				

Table 5: Latency-based evaluation in Task 1. Metric ranking: ERDE30. In bold the best values for each metric are marked.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	MCDI	0	**0.594**	0.605	**0.599**	**0.589**
2	MCDI	1	**0.594**	0.605	**0.599**	**0.589**
3	MCDI	2	**0.594**	0.605	**0.599**	**0.589**
4	HULAT_UC3M	1	0.581	0.589	0.573	0.558
5	ELiRF-UPV	1	0.550	0.564	0.556	0.540
6	ELiRF-UPV	2	0.538	0.543	0.542	0.534
7	ELiRF-UPV	0	0.538	0.535	0.535	0.533
8	PLN_PPM_ISB	2	0.538	0.535	0.534	0.532
9	UC3Mental	1	0.513	0.507	0.506	0.495
10	HULAT_UC3M	2	0.487	0.488	0.488	0.487
11	PLN_PPM_ISB	1	0.569	0.636	0.554	0.484
12	SoloResearch	2	0.506	0.497	0.498	0.475
13	SoloResearch	1	0.500	0.486	0.490	0.455
14	SoloResearch	0	0.494	0.476	0.483	0.446
15	PLN_PPM_ISB	0	0.550	0.632	0.534	0.436
16	UC3Mental	2	0.550	0.632	0.534	0.436
17	Robertuito		0.550	**0.657**	0.533	0.428
18	PUXai	2	0.531	0.576	0.514	0.399
19	PUXai	0	0.531	0.596	0.514	0.390
20	UC3Mental	0	0.506	0.465	0.490	0.385
21	PUXai	1	0.525	0.594	0.507	0.367
22	Roberta Base		0.519	0.259	0.500	0.342
23	VerbaNexAI Lab	0	0.519	0.259	0.500	0.342
24	VerbaNexAI Lab	1	0.519	0.259	0.500	0.342
25	VerbaNexAI Lab	2	0.519	0.259	0.500	0.342
26	I2C-UHU-Rigel	0	0.519	0.259	0.500	0.342
27	I2C-UHU-Rigel	1	0.519	0.259	0.500	0.342
28	I2C-UHU-Rigel	2	0.519	0.259	0.500	0.342
29	HULAT_UC3M	0	0.512	0.258	0.494	0.339
30	NLP-UNED	1	0.244	0.295	0.121	0.165
31	NLP-UNED	2	0.256	0.270	0.126	0.162
32	NLP-UNED	0	0.213	0.230	0.103	0.126

Table 6: Classification-based evaluation in Task 2. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	PLN_PPM_ISB	0	0.316	**0.242**	**2**	**0.990**	0.683
2	PLN_PPM_ISB	1	0.412	0.248	3	0.981	0.680
3	UC3Mental	2	0.415	0.249	3	0.981	0.676
4	VerbaNexAI Lab	0	**0.274**	0.250	**2**	**0.990**	0.677
5	VerbaNexAI Lab	1	**0.274**	0.250	**2**	**0.990**	0.677
6	VerbaNexAI Lab	2	**0.274**	0.250	**2**	**0.990**	0.677
7	I2C-UHU-Rigel	0	0.288	0.250	**2**	**0.990**	0.677
8	I2C-UHU-Rigel	1	0.288	0.250	**2**	**0.990**	0.677
9	I2C-UHU-Rigel	2	0.288	0.250	**2**	**0.990**	0.677
10	Robertuito		0.329	0.252	**2**	**0.990**	**0.685**
11	Roberta Base		0.280	0.256	**2**	**0.990**	0.676
12	HULAT_UC3M	0	0.363	0.256	**2**	**0.990**	0.671
13	PUXai	1	0.577	0.262	5	0.962	0.657
14	HULAT_UC3M	1	0.339	0.271	**2**	**0.990**	0.654
15	PUXai	2	0.370	0.277	3	0.981	0.668
16	UC3Mental	0	0.445	0.283	3	0.981	0.645
17	PUXai	0	0.435	0.299	3	0.981	0.670
18	UC3Mental	1	0.575	0.334	7	0.942	0.556
19	PLN_PPM_ISB	2	0.454	0.340	4	0.976	0.570
20	MCDI	0	0.381	0.343	**2**	**0.990**	0.540
21	MCDI	1	0.381	0.343	**2**	**0.990**	0.540
22	MCDI	2	0.381	0.343	**2**	**0.990**	0.540
23	ELiRF-UPV	2	0.600	0.394	14	0.876	0.432
24	ELiRF-UPV	1	0.579	0.402	14	0.876	0.412
25	ELiRF-UPV	0	0.649	0.410	21	0.810	0.470
26	SoloResearch	0	0.639	0.424	11	0.904	0.550
27	SoloResearch	1	0.674	0.458	20	0.820	0.501
28	HULAT_UC3M	2	0.624	0.472	20	0.82	0.4
29	SoloResearch	2	0.691	0.487	26	0.769	0.464
30	NLP-UNED	2	0.762	0.655	6	0.952	0.399
31	NLP-UNED	0	0.756	0.688	4	0.971	0.399
32	NLP-UNED	1	0.780	0.693	6	0.957	0.377

Table 7: Latency-based evaluation in Task 2. Metric ranking: ERDE30. In bold the best values for each metric are marked.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	PLN_PPM_ISB	1	**0.938**	**0.952**	**0.915**	**0.927**
2	PLN_PPM_ISB	0	0.931	0.923	0.906	0.912
3	ELiRF-UPV	1	0.913	0.929	0.877	0.887
4	ELiRF-UPV	2	0.888	0.904	0.862	0.873
5	SoloResearch	0	0.900	0.928	0.850	0.856
6	SoloResearch	1	0.900	0.928	0.850	0.856
7	SoloResearch	2	0.900	0.933	0.846	0.850
8	ELiRF-UPV	0	0.881	0.895	0.829	0.832
9	PLN_PPM_ISB	2	0.875	0.902	0.823	0.825
10	PUXai	2	0.869	0.902	0.821	0.822
11	Robertuito		0.844	0.830	0.808	0.813
12	Roberta Base		0.869	0.896	0.810	0.804
13	VerbaNexAI Lab	0	0.813	0.846	0.769	0.780
14	VerbaNexAI Lab	1	0.813	0.846	0.769	0.780
15	VerbaNexAI Lab	2	0.813	0.846	0.769	0.780
16	UC3Mental	2	0.863	0.917	0.794	0.778
17	UC3Mental	0	0.863	0.917	0.794	0.778
18	UC3Mental	1	0.863	0.917	0.794	0.778
19	PUXai	0	0.806	0.889	0.736	0.722
20	MCDI	0	0.838	0.900	0.756	0.721
21	MCDI	1	0.838	0.900	0.756	0.721
22	MCDI	2	0.838	0.900	0.756	0.721
23	PUXai	1	0.806	0.859	0.733	0.713
24	NLP-UNED	1	0.731	0.566	0.654	0.590
25	NLP-UNED	2	0.706	0.576	0.635	0.577
26	NLP-UNED	0	0.663	0.515	0.594	0.529
27	I2C-UHU-Rigel	0	0.450	0.433	0.427	0.361
28	I2C-UHU-Rigel	1	0.450	0.433	0.427	0.361
29	I2C-UHU-Rigel	2	0.450	0.433	0.427	0.361
30	HULAT_UC3M	0	0.244	0.184	0.242	0.192
31	HULAT_UC3M	1	0.244	0.184	0.242	0.192
32	HULAT_UC3M	2	0.244	0.184	0.242	0.192

Table 8: Classification-based evaluation for multi-class in Task 2. Metric ranking: Macro-F1. In bold the best values for each metric are marked.