I2C-UHU-Rigel at MentalRiskES 2025: Detection of Gambling Disorder Risk in Spanish using Transformer-Based Models

Antonio L. García Moreno[1], Jacinto Mata Vázquez[2] and Victoria Pachón Álvarez[3]

I2C Research Group, University of Huelva, Spain

Abstract

This paper presents the approaches proposed by the I2C Group to address the MentalRiskES task on early detection of mental disorder risks in Spanish, as part of IberLEF 2025. Our proposal involves developing and fine-tuning various transformer-based models to handle two subtasks: (i) a binary classification task for identifying gambling disorder risk, in which we determine whether a user is at high risk (label = 1) or low risk (label = 0); and (ii) a multiclass classification task for identifying the specific type of addiction. In the latter case, assuming that all users are at risk, the system assigns each to one of four categories: Betting, Online Gaming, Trading, or Lootboxes, based on their message history. Our core methodology relies on fine-tuning pre-trained transformer models (e.g., BETO, XLM-RoBERTa, and DistilBERT) on annotated message sets. For the binary task, we first generate message-level predictions, which are then aggregated per user using soft voting and thresholding to compute robust user-level risk scores. For the multiclass addiction-type task, we address severe class imbalance through targeted data augmentation techniques (including back-translation and synonym replacement) as well as loss-weighted fine-tuning strategies. The final system achieved a Macro-F1 score of 0.551 in Task 1, ranking 3rd overall, and a Macro F1-Score of 0.342 in Task 2, ranking 27th.

Keywords

Mental Health, Gambling, Early Detection, Transformer, Deep Learning, Natural Language Processing

1. Introduction

Gambling disorder is recognized by the World Health Organization as a behavioral addiction with significant global impact [1].

In Spain, the 2024 EDADES survey reports that 53.8% of residents aged 15–64 engaged in some form of gambling during the past year, predominantly lotteries and sports betting [2]. Of these, approximately 1.4% exhibit signs of problematic gambling, with 0.4% meeting full disorder criteria and an additional 1% classified as moderate risk [3].

Digital transformation has reshaped gambling practices. Online formats emerging as the fastest-growing segment despite stable or declining participation in traditional formats [2, 3]. Sports betting dominates online gambling and serves as an entry point for younger individuals. Emerging loot box mechanics in video games mirror wagering behaviors and show a strong correlation with problem gambling indicators among adolescents [4].

In this shared task, we present our approach, which leverages advances in transformer-based language models to develop and evaluate systems for early detection of gambling disorder risk in social media streams, addressing both risk classification and subtype identification. For training data preparation, each user was individually labeled on each task. Message-level predictions were then aggregated per user via soft voting to obtain robust user-level scores. We addressed class imbalance in the multiclass task through back-translation data augmentation, and conducted a Bayesian hyperparameter sweep over diverse parameters. Additionally, we experimented with chunk-based segmentation to effectively handle message history.

The remainder of this paper is organized as follows. Section 2 reviews related work on transformer-based approaches on behavioral addiction detection. Section 3 describes the dataset and task definitions.

IberLEF 2025, September 2025, Zaragoza, Spain

antonioluis.garcia420@alu.uhu.es (A. L. G. Moreno); mata@uhu.es (J. M. Vázquez); vpachon@dti.uhu.es (V. P. Álvarez)

Section 4 details our modeling and implementation strategies. Section 5 presents experimental results and analysis. Section 6 discussed the findings and limitations. Finally, Section 7 concludes and outlines future directions.

2. Related Work

Previous research on the automatic detection of behavioral addictions and mental health conditions has employed deep neural networks and hand-crafted features. Transformer-based models, such as BERT and RoBERTa, have achieved state-of-the-art results in binary classification tasks, such as depression screening [5] and substance use detection [6]. Multilingual models demonstrate robustness across languages, but often underperform compared to language-specific variants in specialized domains [7].

Despite extensive research on sentiment analysis and content moderation in social media [8], comparatively little work has addressed the automatic detection of behavioral addictions such as gambling disorder. Most studies have focused on more prevalent or socially visible mental health concerns, leaving behavioral addictions underexplored in NLP and social media analysis contexts. This gap highlights the need for further research targeting the early identification and monitoring of these conditions, especially given their growing prevalence and impact [2].

3. Data

Organizers provided per-user JSON files—each containing an average of 64 messages—comprising timestamped content from Telegram and Twitch. Additionally, a separate file was provided containing the label for each user. After downloading and merging the data, we obtained 350 users, with the following stratified splits: 280 users (80%) for training and 70 users (20%) for testing in both Task1 and Task2.

- Task 1: 280 users for training and 70 for testing.
- Task 2: 280 users for training and 70 for testing.

Task 1 dataset consists of six columns: *id message*, *message*, *date*, *platform*, *user* and *label* (see Table 1).

Table 1
Example of training dataset for Task 1

id_message	message	date	platform	user	label
21303110347	al que le toca es a gaben	2020-08-13 08:23:09+01:00	Twitch	user28847	0
99473399904	ya no abres cajas con el rabo	2020-08-13 08:23:12+01:00	Twitch	user28847	0
3937262479	bien de culo	2020-08-13 08:23:22+01:00	Twitch	user28847	0

Task 2 is also composed of six columns: *id_message*, *message*, *date*, *platform*, *user* and *label* (see Table 2).

Table 2
Example of training dataset for Task 2

id_message	message	date	platform	user	label
66765627559	por gorda	2023-02-23 00:15:08+01:00	Twitch	user17509	onlinegaming
28838294712	en wh esto si es posible xr	2019-07-28 13:13:46+01:00	Telegram	user4180	betting
64232095759	btc long será??	2021-03-27 17:39:15+01:00	Telegram	user7529	trading

After constructing the datasets, they were split into three subsets: training, validation, and test. Eighty percent of the data were allocated to the training and validation sets, and 20% to the test set.

Within the train set, 25% was allocated for validation. Tables 3–5 detail the class distributions for each task across these subsets.

Table 3
Distribution of users and messages per data split

Task	Train		Valid		Test	
	Users	Messages	Users	Messages	Users	Messages
Task 1	210	13 666	70	4 384	70	4 441
Task 2	210	13 487	70	4 482	70	4 522

Table 4
Class distribution of datasets in Task 1

Label	**Train dataset**		**Valid dataset**		**Test dataset**	
	Users	**Messages**	**Users**	**Messages**	**Users**	**Messages**
0	106	7 162	36	2 306	36	1 971
1	104	6 504	34	2 078	34	2 470

Table 5
Class distribution of datasets in Task 2

Label	**Train dataset**		**Valid dataset**		**Test dataset**	
	Users	**Messages**	**Users**	**Messages**	**Users**	**Messages**
betting	51	5 628	17	1 985	17	1 953
lootboxes	16	134	5	40	5	41
onlinegaming	62	1 615	21	551	21	537
trading	81	6 110	27	1 906	27	1 991

As it was detailed previously, this paper is focused on both tasks. Task 1 is a binary classification in which must be detected if the user is at low or high risk. Labels will be 0 for "*low risk*" or 1 for "*high risk*". Task 2 is a multiclass classification. The model must predict one of the four presented classes ("*betting*", "*lootboxes*", "*onlinegaming*", "*trading*").

4. Methodology and experiments

4.1. Task 1: Risk Detection of Gambling Disorders

For Task 1, we fine-tuned three transformer-based models available on Hugging Face [9]: BETO (bert-base-spanish-wwm-cased) [10], XLM-RoBERTa (xlm-roberta-base) [7] , and DistilBERT (distilbert-base-uncased) [11]. To prepare the data, we applied several distinct preprocessing methods, each paired with different chunk sizes (128, 256, and 512 tokens). As part of the experimentation, three different variants were applied:

- **Variant A**: Applies basic normalization using regex: lowercasing; removal of URLs (e.g., http, bit.ly), placeholders ([link], [url]), retweet markers (RT @user), usernames (@user), generic [user] tokens, and hashtags. Finally, all emojis are removed using a comprehensive Unicode emoji pattern.
- **Variant B**: Extends Variant A by preserving hashtag content (removing only the '#' symbol) and remapping selected emojis to retain emotional cues.

- **Variant C**: Incorporates Spanish spelling correction via Speller, expands common abbreviations (e.g., ’q’ → ’que’, ’xq’ → ’porque’), and then applies the same regex-based cleaning and emoji remapping as in VariantB.

After normalization, texts were tokenized and segmented into non-overlapping chunks, preserving the original timestamp order to retain contextual coherence. While using default hyperparameters, to identify the configurations that maximized contextual information while minimizing noise, we conducted experiments with all preprocessing–chunk size combinations, as shown in Table 6.

Table 6
Accuracy and F1-Score results

Model	Accuracy	F1-Score
BETO_128_A	0.53	0.54
BETO_128_B	0.60	0.63
BETO_128_C	0.61	0.60
BETO_256_A	0.59	0.59
BETO_256_B	0.61	0.64
BETO_256_C	0.62	0.62
BETO_512_A	0.62	0.62
BETO_512_B	**0.66**	**0.66**
BETO_512_C	0.60	0.63
XLM-RoBERTa_128_A	0.52	0.53
XLM-RoBERTa_128_B	0.57	0.58
XLM-RoBERTa_128_C	0.65	0.64
XLM-RoBERTa_256_A	0.58	0.58
XLM-RoBERTa_256_B	0.57	0.58
XLM-RoBERTa_256_C	0.59	0.59
XLM-RoBERTa_512_A	**0.69**	**0.68**
XLM-RoBERTa_512_B	0.64	0.66
XLM-RoBERTa_512_C	0.64	0.63
DistilBERT_128_A	0.45	0.47
DistilBERT_128_B	0.48	0.50
DistilBERT_128_C	0.49	0.51
DistilBERT_256_A	**0.61**	**0.61**
DistilBERT_256_B	0.52	0.51
DistilBERT_256_C	0.52	0.53
DistilBERT_512_A	0.49	0.50
DistilBERT_512_B	0.61	0.60
DistilBERT_512_C	0.47	0.49

After the training process and generating predictions for the individual messages, we employed a user prediction strategy for the task. The user will be labeled as high risk or not based on a threshold approach. To determine the minimum threshold value, we conducted an additional experiment in which we applied various thresholds to identify when the metrics peaked, thereby ensuring the model’s optimal performance. By using this method, we ensure that each user’s label corresponds to the most common classification inferred from the predicted labels of their messages, obtaining a representative label.

We then performed a Bayesian hyperparameter search over learning rate, weight decay, and batch sizes using Weights & Biases [12], as detailed in Table 7.

Table 7
Bayesian search hyperparameters

Hyperparameters	**Values**
learning_rate	5e-5, 3e-5, 2e-5, 1e-5
weight_decay	0.0, 0.01, 0.1
per_device_train_batch_size	8, 16, 32
per_device_eval_batch_size	8, 16, 32

The best hyperparameter configurations for each model are presented in Table 8, and the corresponding test results are summarized in Table 9.

Table 8
Best configuration of hyperparameters

Model	**learning_rate**	**weight_decay**	**train batch_size**	**eval batch_size**
bert-base-spanish-wwm-cased	3e-5	0.1	16	16
xlm-roberta-base	1e-5	0.1	8	8
distilbert-base-uncased	1e-5	0.1	16	16

Table 9
Model results F1-Score and accuracy

Model	**F1-Score**	**Accuracy**
bert-base-spanish-wwm-cased	0.69	0.69
xlm-roberta-base	**0.69**	**0.70**
distilbert-base-uncased	0.61	0.61

Figure 1 illustrates a summary of the methodology implemented for Task 1.

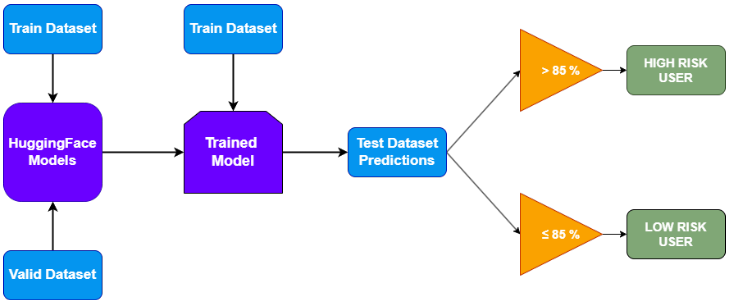

Figure 1: Methodology for Task 1

While all three architectures demonstrate solid performance in early gambling-risk detection, XLM-RoBERTa emerges as the clear frontrunner with A F1-Score of 0.70 (vs. 0.69 for the BETO), suggesting that its multilingual pretraining better captures the nuanced language of at-risk posts. DistilBERT's lower F1 of 0.61 indicates that lightweight models may underfit this domain without additional distillation or adapter strategies.

To further maximize model performance in the binary classification of high-risk users, we systematically evaluated the impact of varying the decision threshold applied to the XLM-RoBERTa model, which had previously demonstrated the best results in our experiments. Specifically, three threshold values (0.30, 0.50, and 0.75) were tested to determine the optimal balance between sensitivity and specificity. The results are summarized in Table 10.

Table 10
Model Performance at Different Thresholds

Threshold	Accuracy	F1-Score
0.30	0.49	0.33
0.50	0.69	0.70
0.75	0.76	0.73

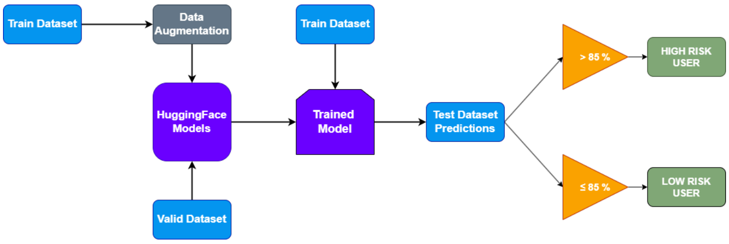

Figure 2: Methodology for Task 2

As observed, increasing the threshold leads to a reduction in false positives, with the threshold of 0.75 achieving the highest accuracy and F1-Score. This suggests that predicted probabilities may be biased toward intermediate values, necessitating a higher threshold to confidently identify positive cases while minimizing the inclusion of misclassified instances. However, from a clinical perspective, it is often preferable to tolerate a higher rate of false positives rather than risk missing users who are genuinely at risk. This trade-off underscores the importance of aligning threshold selection with the intended application context, balancing statistical performance with real-world implications

4.2. Task 2: Type of addiction

To address Task 2, we directly built upon the setup used in Task 1, reusing the same model architectures and preprocessing pipelines.. This consistency ensured that any performance differences truly reflect the new task's demands rather than changes in model tuning.

Initial results did not meet expectations, prompting us to explore additional strategies to mitigate class imbalance. First, the pronounced class imbalance made it necessary to augment the training set—without this boost, minority labels simply weren't getting learned. In fact, we eventually narrowed the task to three categories after finding that the models struggled to recognize the "lootboxes" class—possibly because loot box discussions exhibit similar linguistic patterns to trading and online gaming, making it challenging for the model to distinguish them clearly. Back-translation was performed using the NLLB-200-distilled-600M model from Facebook, available on Hugging Face, translating Spanish examples to Japanese and Arabic and back to Spanish, thereby augmenting underrepresented classes.

After augmentation, the class distributions changed as shown in Table 11.

Table 11
Train dataset distribution before and after data augmentation was applied

	trading	betting	onlinegaming	lootboxes
Before Data Augmentation	6 110	5 628	1 615	134
After Data Augmentation	5 976	5 621	4 532	268

To determine a user's final label, we aggregated all predicted message-level categories and selected the most frequent one—a majority vote strategy that helps mitigate occasional misclassifications at the message level. By letting the most frequently predicted class prevail, we ensure that our final user-level assignments reflect the dominant pattern in their communications.

Figure 2 illustrates methodology used in Task 2.

As summarized in Table 12, data augmentation improved performance for ’onlinegaming’ but failed to boost recognition for ’lootboxes’. This indicates that simply increasing the quantity of minority-class samples wasn’t sufficient, loot box references continue to slip through unnoticed, highlighting the need for more targeted feature engineering or alternative augmentation techniques to better capture this label.

Table 12
F1-Score per class and Macro-F1 for each model

Model	F1 Trading	F1 Betting	F1 Onlinegaming	F1 Lootboxes	Macro-F1
bert-base-spanish-wwm-cased	0.63	0.52	0.36	0.00	0.38
xlm-roberta-base	0.69	0.56	0.45	0.02	0.43
distilbert-base-uncased	0.53	0.42	0.26	0.00	0.30

Table 12’s performance gains line up clearly with how we rebalanced the training set (see distribution table above): by boosting the “onlinegaming” examples from 1,615 to 4,532 and doubling “lootboxes” from 134 to 268—while only slightly trimming “trading” and “betting”—we aimed to give the minority classes a fair shot. Indeed, the F1 for online gaming climbs to 0.45 under XLM-RoBERTa, suggesting that sheer volume does help the model learn those patterns. However, “lootboxes” stubbornly remains at—or near—zero F1 across all architectures, dragging our Macro-F1 down to the low 0.3–0.4 range. In other words, even after equalizing counts, the model still fails to capture the specific language of loot-box talk.

As the “lootboxes” category continues to resist all augmentation efforts—its F1-Score stubbornly hovering at zero, we then investigated excluding the ’lootboxes’ class entirely, but overall performance declined, indicating that even underrepresented labels carry valuable contextual information (see Table 12). By reducing the label space down to just trading, betting and online gaming, we aim to eliminate the noise introduced by an almost-unlearnable category. This streamlined three-class configuration should allow the model to concentrate fully on the remaining, well-represented behaviors and may yield cleaner decision boundaries and higher overall performance. Table 13 shows results obtained by each model after removing "lootboxes" class.

Table 13
F1-Scores and Macro-F1 after removing lootboxes

Model	F1 Trading	F1 Betting	F1 Onlinegaming	Macro-F1
bert-base-spanish-wwm-cased	0.94	0.92	1.00	0.95
xlm-roberta-base	0.96	0.69	0.32	0.66
distilbert-base-uncased	0.78	0.77	0.65	0.73

5. Results

During the evaluation phase, all three models were employed for Task 1, while only BETO was used for Task 2. Our system performed better than expected in Task 1, achieving 3rd place with XLM-RoBERTa, as shown in Table 15.

In contrast, performance in Task 2 was significantly lower, with our system ranking 27th, as shown in Table 14.

Table 14
Competition results for Task 1

Rank	Team	Accuracy	Macro F1	Micro F1	ERDE5	ERDE30	Latency-weighted F1
1	UNSL	0.569	0.567	0.569	0.639	0.389	0.506
2	UNSL	0.581	0.563	0.581	0.515	0.284	0.628
3	**I2C-UHU-Rigel**	**0.556**	**0.551**	**0.556**	**0.600**	**0.432**	**0.496**

Table 15
Competition results for Task 2

Rank	Team	Accuracy	Macro F1	Micro F1	ERDE5	ERDE30	Latency-weighted F1
1	PLN_PPM_ISB	0.569	0.484	0.569	0.412	0.248	0.680
12	Roberta Base	0.519	0.342	0.519	0.280	0.256	0.676
27	**I2C-UHU-Rigel**	**0.519**	**0.342**	**0.519**	**0.288**	**0.250**	**0.677**
32	HULAT_UC_3M	0.487	0.487	0.487	0.624	0.472	0.400

Limiting training to only three categories—trading, betting, and online gaming—and excluding 'lootboxes' did not improve performance, suggesting that even sparsely populated classes contribute essential signals.

6. Error Analysis

Error analysis is a crucial stage in any machine learning project, as it provides insights into model limitations and highlights instances where the system fails to perform optimally. This enables the exploration of alternative approaches and the proposal of targeted improvements.

In this section, we examine the errors made during the evaluation phase by the best-performing model for each task: XLM-RoBERTa for Task 1 and BETO for Task 2. The analysis focuses on identifying common characteristics among misclassified examples to extract detailed conclusions.

6.1. Task 1: Gambling Risk Detection

For Task 1, we observed a high rate of false positives, with more than half of users classified as high risk. This suggests that any language related to finance or gambling is often treated as an indicator of risk.

At the message level, results show a bias toward high-risk detection—a conservative approach that prioritizes not missing real cases of problematic behavior, even at the expense of over-predicting risk. Figure 3 presents the confusion matrix for message-level evaluation in Task 1.

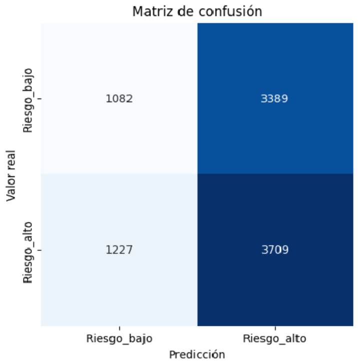

Figure 3: Confusion matrix at message level during evaluation for Task 1

Error patterns among false positives often involve ambiguous terms that lead to incorrect predictions. Table 16 provides representative examples, their real context, and the underlying cause of the error.

Table 16
Representative instances of problematic patterns in false positives

Example	True Context	Source of Error
"Acabo de ver las slots *GREAT RHINO* o *MEDUSA megaways*"	Passive observation without active engagement	Misinterpretation of game-related terminology
"Y si es a la inversa iríamos hasta los 30 mil"	Speculation conveyed with humor	Literal recognition of numerical expressions without contextual analysis
"Eth será que sube a 17"	Neutral commentary on cryptocurrency performance	Bias toward cryptocurrency token mentions

For false negatives, errors are often caused by the absence of explicit financial terminology or the use of informal vocabulary that the model fails to recognize (see Table 17).

Example	Problematic Context	Source of Error
"Me escribió un mensaje el *unknown* diciendo que mire en solicitud de mensaje de discofd"	Messages pertaining indirectly to trading/gambling	Absence of explicit financial terminology
"yo hice así, ahora me cebo y me saco 60"	"me cebo" = compulsive behavior	Unrecognized informal vocabulary

Table 17
Representative instances of undetected patterns in false negatives

Another area of interest has been to examine how message length affects the error rate, in order to determine whether a consistent pattern can be observed. Figure 4 illustrates the evolution of the error rate as a function of text length.

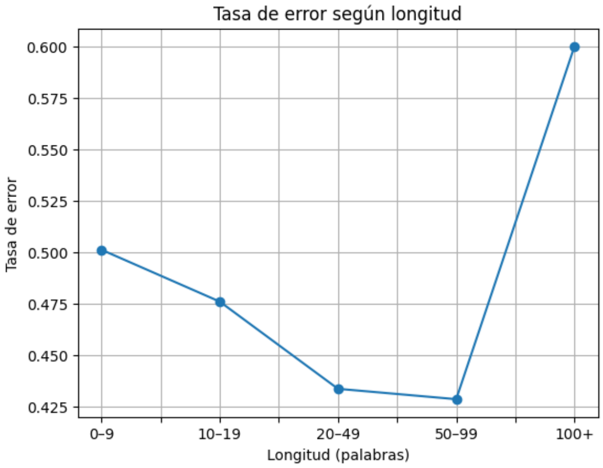

Figure 4: Error rate evolution based on message length in Task 1

Long messages—comprising the smallest segment of the evaluation set—exhibit the highest error rate. However, owing to their underrepresentation in the dataset, it cannot be conclusively determined that length itself is the causal factor; indeed, the observed trend suggests that as message length increases (and thus provides greater context), the model's error rate actually decreases.

Finally, to conclude the analysis of Task 1, we examined the most frequent terms in the misclassified examples to determine whether they reveal lexical features that have contributed to bias. Table 18 presents the ten most frequent terms alongside a possible explanation of the contextual issue each term introduces.

Table 18
Frequent tokens and their possible problematic contexts

Token	Frequency	Problematic Context
btc	189	Any mention of Bitcoin is construed as indicating high risk.
short	76	Failure to distinguish between a trading strategy and a loss.
bien	68	Neutral expressions are misinterpreted as having sentiment.
creo	57	Expressions of uncertainty are interpreted as anxiety.
mas	55	“Increase” versus a neutral quantitative statement.
unknown	55	References to users potentially involved in betting.
long	50	Technical position terminology conflated with behavioral context.
ahora	47	Temporal urgency inferred from simple time reference.
solo	46	Expressions of isolation or mere quantitative specification.

The model interprets any reference to financial terminology, as well as neutral terms, as indicators of risk-related behavior; in other words, it fails to distinguish an analytical discussion of financial matters from the compulsive behaviors characteristic of this disorder.

6.2. Task 2: Type of addiction

For Task 2, there is a tendency to cluster predictions into the betting and trading classes, reflecting both the imbalance of the provided dataset and the model’s limitations in distinguishing between classes with similar contexts. Figure 5 presents the confusion matrix obtained during the evaluation phase of Task 2.

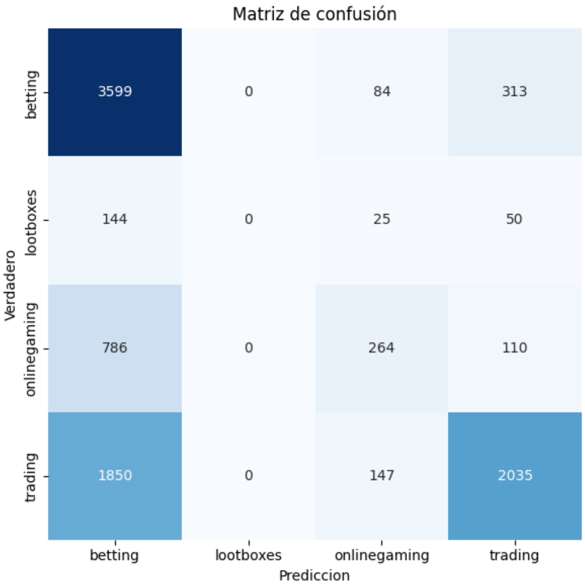

Figure 5: Confusion matrix at message level during evaluation for Task 2

With regard to the dominant patterns in the misclassified examples, we focus specifically on those instances that belonged to the betting class but were classified as trading. As Table 19 illustrates, there is considerable overlap in the terminology used by both classes (trading and betting), causing the model to gravitate toward the class that was more heavily represented during training.

Table 19
Representative cases with real context and error causes

Example	True Context	Source of Error
"Muchas ofertas las que tuvieron, casi todo a mitad de precio"	Market analysis of cryptocurrency/trading	The terms "ofertas" and "precio" are misinterpreted as betting promotions.
"Hola!!. Saben para qué sirve el nft de oro que te da Binance por acertar el resultado exacto del partido del mundial?"	Cryptocurrency trading (Binance)	The phrase "acertar resultado del partido" combined with "mundial" is interpreted as sports betting.

Table 20
Performance and energy consumption metrics for XLM-RoBRTa run

Metric	Mean Value	Std. Dev.	Description
Duration (hours)	0.0105	0.00995	Training time per run
CO_2e Emissions (kg)	0.00419	0.00499	Estimated emissions per run
CPU Energy (kWh)	0.00292	0.00292	CPU energy consumption
GPU Energy (kWh)	0.00376	0.00376	GPU energy consumption
RAM Energy (kWh)	0.000627	0.000627	RAM energy consumption
Total Energy (kWh)	0.00737	0.00737	Total energy consumed per run

With respect to message length, it is evident that the error rate exhibits a linear relationship with input length, thereby reinforcing the notion that concatenating messages is necessary to capture as much context as possible within the constraints of transformer-based models (i.e., a 512-token maximum). Figure 6 depicts the evolution of the error rate as a function of message length.

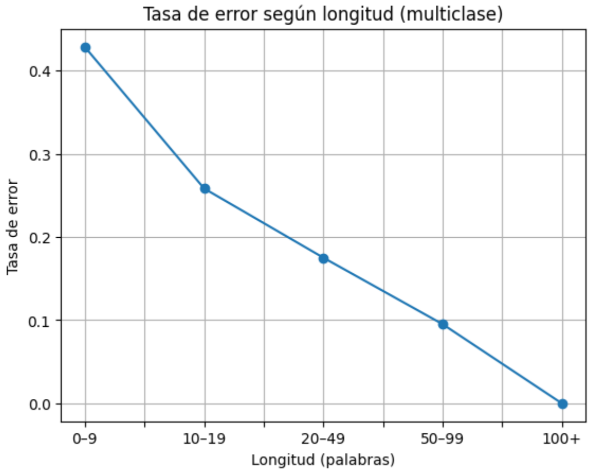

Figure 6: Error rate evolution based on message length in Task 2

Finally, regarding the most frequent words, no significant differences from Task 1 are observed, and therefore they do not contribute any relevant information to the analysis.

7. Enviromental impact

The environmental impact of model training was quantified using the CodeCarbon tool, in line with established guidelines for NLP research and recent recommendations for climate-aware reporting. All experiments were conducted on an Intel(R) Xeon(R) CPU @ 2.20GHz and a Tesla T4 GPU, with metrics averaged across repeated runs. See Table 20 for performance and energy consumption results.

The reported emissions are situated at the lower end of the spectrum for transformer-based NLP experiments [], which is consistent with findings that emissions can vary widely depending on hardware efficiency (1x Tesla T4), dataset size, and training duration. The relatively low emissions are attributed to the use of energy-efficient hardware, short training times, and resource-conscious experimental design—factors recognized as best practices for sustainable AI research. By quantifying and reporting these metrics, this work aligns with the growing movement towards responsible and climate-aware NLP research.

8. Conclusions

In this paper, we presented the I2C-UHU-Rigel approach to the MentalRiskES [13] shared task at IberLEF 2025 [14], which focused on early detection of gambling disorder risk (Task 1) and multiclass classification of addiction modality (Task 2) in Spanish. Our core methodology—based on fine-tuning BETO, XLM-RoBERTa, and DistilBERT on message-level data, combined with careful preprocessing, Bayesian hyperparameter search using Weights & Biases, threshold optimization, and majority-vote aggregation—yielded strong performance in Task 1, with XLM-RoBERTa achieving a Macro F1-Score of 0.551 and ranking 3rd in the competition.

Due to class imbalance, Task 2 proved to be more challenging; despite using back-translation and loss-weighted fine-tuning, the 'lootboxes' label remained largely unrecognized, resulting in a Macro F1-Score of 0.342 (27th place). Follow-up experiments showed that reducing the label space by removing the problematic lootboxes class dramatically improves performance on the remaining categories (e.g. BETO reaches a Macro-F1 of 0.95), highlighting the trade-off between label granularity and model learnability.

Future work will explore ensemble methods to combine complementary strengths of our transformer models, domain-adaptive pretraining on large-scale gambling-related corpora, and more advanced augmentation or adversarial example generation methods targeting underrepresented categories.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to assist with language grammar/spelling and minor paraphrasing. The authors reviewed and edited all AI-assisted text and take full responsibility for the final manuscript.

References

[1] World Health Organization, Icd-11 mms: 6c50 gambling disorder, International Classification of Diseases for Mortality and Morbidity Statistics, 11th Revision, v2025-01, 2025. URL: https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/1448597234.

[2] Observatorio Español de las Drogas y las Adicciones (OEDA), Informe sobre adicciones comportamentales 2024, Technical Report, Delegación del Gobierno para el Plan Nacional sobre Drogas (DGPNSD), 2024. URL: https://pnsd.sanidad.gob.es/profesionales/sistemasInformacion/boletines/edades/, "El 53,8% de la población de 15 a 64 años ha jugado a juegos de azar (presencial y online)".

[3] Ministerio de Sanidad, España, Informe sobre adicciones comportamentales y otros trastornos adictivos, Technical Report, Plan Nacional sobre Drogas, Gobierno de España, 2024. URL: https://www.mscbs.gob.es/ciudadanos/enfLesiones/enf_mentales/, "La prevalencia del juego problemático se redujo un 46% desde 2018, afectando al 1,4% de la población".

[4] G. Brooks, L. Clark, A longitudinal replication study testing migration from video game loot boxes to gambling, PubMed Central (2025). "El gasto en loot boxes predice iniciación en apuestas convencionales (OR = 1.32)".

[5] J. Novikova, K. Shkaruta, et al., DECK: Behavioral tests to improve interpretability and generalizability of bert models detecting depression from text, arXiv preprint arXiv:2209.05286 (2022).

[6] A. Sarker, et al., Social media-based monitoring for substance use, in: Digital Ethology: Human Behavior in Geospatial Context, MIT Press, 2019. "Estimaciones derivadas de Twitter/X sobre el uso de opioides se correlacionaron con muertes por sobredosis".

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[8] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, arXiv preprint arXiv:1703.04009 (2017).

[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771, 2019.

[10] J. Cañete, C. Cardellino, et al., spanish-bert: Spanish pre-trained bert model, arXiv preprint arXiv:2004.06205, 2020.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[12] Weights & Biases, Weights & biases, https://wandb.ai/, 2025.

[13] A. M. Mármol-Romero, P. Álvarez Ojeda, 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 2025: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 75 (2025).

[14] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, 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, 2025.

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