PUXai at MentalRiskES 2025: RoBERTuito with Bidirectional LSTM for Early Gambling Disorder Detection

Nguyen Xuan Phuc[1,2,*], Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

The increasing prevalence of gambling disorders underscores the importance of early detection, a challenge addressed by the MentalRiskES 2025 shared task. This paper presents our solution for two tasks: binary classification for detecting gambling disorder risk (Task 1) and multiclass classification for identifying addiction types (Task 2). Using pretrained Spanish language models, task-specific data augmentations, and robust optimization strategies, we achieved the highest Macro F1 scores in both tasks. For Task 1, we use RoBERTuito with an LSTM classifier and GroupDRO to detect gambling disorder risk. For Task 2, we employ roberta-base-bne with back-translation and Optuna-based hyperparameter tuning to identify addiction types. Data preprocessing and augmentation techniques were employed to mitigate issues of data scarcity and imbalance. Our solution ranks 21st for Task 1 and 10th for Task 2 in the MentalRiskES 2025 shared task.

Keywords

Early Risk Detection, Gambling Disorders, RoBERTuito, roberta-base-bne, Data Augmentation, GroupDRO, Optuna

1. Introduction

The rise of online gambling platforms has led to a surge in gambling disorders, characterized by compulsive betting behaviors that cause significant financial, psychological, and social harm [1]. Early detection of these disorders is critical to enable timely interventions, such as counseling or restricted platform access, which can mitigate severe consequences. However, identifying at-risk individuals in real-time using social media data, such as messages from platforms like Telegram and Twitch, presents substantial challenges due to the informal, sequential, and often ambiguous nature of the data [2, 3].

The primary problem is to detect gambling disorder risk and classify specific addiction types (e.g., betting, lootboxes) from user-generated messages. For risk detection, the goal is to distinguish high-risk from low-risk individuals based on sequential message patterns that may indicate compulsive behavior. For addiction type classification, the task is to identify the specific gambling-related addiction driving the behavior. A key difficulty is the sequential nature of real-world data: messages are generated incrementally over time, requiring models to make predictions with partial information to achieve early detection. In contrast, training data often consists of complete message sequences provided at once, creating a mismatch that complicates model training. This static training data makes it challenging to simulate real-time decision-making, as models may overfit to full sequences rather than learning to detect early risk signals.

Additional challenges exacerbate the problem. The informal language in social media messages, replete with emojis, slang, and typographical errors, hinders semantic understanding. Data scarcity, particularly for minority classes like lootbox-related addictions, limits model generalization.

IberLEF 2025, September 2025, Zaragoza, Spain
*Corresponding author.
23521213@gm.uit.edu.vn (N. X. Phuc); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0009-0007-3757-2570 (N. X. Phuc); 0000-0001-8340-1405 (D. V. Thin)

Class imbalance further complicates training, as high-risk or specific addiction types are underrepresented. Moreover, the need for early detection demands models that prioritize rapid identification of risk signals, often at the cost of higher false positives, as delayed predictions reduce intervention efficacy.

To address these challenges, transformer-based large language models pretrained on Spanish corpora, such as RoBERTuito and roberta-base-bne, offer robust solutions for processing informal text. RoBERTuito, fine-tuned for sentiment analysis, excels at generating embeddings that capture nuanced meanings in Spanish social media messages [4]. Pairing RoBERTuito with a bidirectional Long Short-Term Memory (LSTM) network is particularly effective for risk detection, as the LSTM models temporal dependencies in message sequences, enabling early identification of risk patterns [5]. For addiction type classification, roberta-base-bne, pretrained on diverse Spanish texts, supports multiclass classification by leveraging contextual embeddings [6]. Data augmentation techniques, such as negative sample subsequencing and back-translation, mitigate data scarcity and imbalance by generating diverse training samples [7, 8]. Optimization strategies like Group Distributionally Robust Optimization (GroupDRO) [9] and hyperparameter tuning with Optuna [10] enhance model robustness by focusing on harder samples and fine-tuning performance.

This paper presents our solutions for early gambling disorder detection, emphasizing a RoBERTuito-based bidirectional LSTM approach for risk detection and roberta-base-bne for addiction type classification. Section 3 describes the tasks, datasets. Section 4 details our preprocessing, augmentation, and training methods. Section 5 describe evaluation metrics. Section 6 analyzes our results, and Section 8 provides conclusions and future directions.

2. Related Work

Gambling disorder, characterized by compulsive gambling behaviors with severe psychological and financial consequences, has received limited attention in natural language processing (NLP) for mental health risk detection. While prior studies have explored mental health issues using social media data, few have targeted gambling disorders, particularly on informal and noisy platforms like Telegram and Twitch.

Parfenova et al.[11] proposed a sequential model combining GRU and LSTM, enhanced with temporal and emotional features, to detect pathological gambling from English Reddit posts . Their approach, leveraging structured and relatively clean Reddit data, achieved high performance but may not generalize to noisier, real-time platforms like Telegram or Twitch. Similarly, studies on Twitch have focused on sentiment analysis of gambling-related content rather than risk detection [12]. The scarcity of labeled datasets for gambling disorders, coupled with the multilingual and informal nature of these platforms, poses significant challenges for developing robust, generalizable models.

Our work addresses these gaps by adapting transformer models (RoBERTuito and roberta-base-bne) and bidirectional LSTMs for gambling disorder detection on Telegram and Twitch data. Unlike prior studies, we tackle severe class imbalance using back-translation augmentation and enhance early detection through GroupDRO and Optuna-based hyperparameter tuning. By focusing on gambling-specific linguistic patterns (e.g., references to betting platforms or lootboxes), our approach better captures the behavioral cues of affected users, offering a novel contribution to mental health risk detection in noisy, real-time environments.

3. Tasks

The MentalRiskES 2025 [13] shared task, part of IberLEF 2025 [14], focuses on the early detection of mental disorders, specifically gambling disorders, based on user comments from Telegram and Twitch.

The tasks aim to identify whether users suffer from a gambling disorder and the specific type of addiction influencing their mental health, using a stream of messages.

3.1. Overview

All tasks involve detecting gambling disorders in users based on their comments posted on Telegram and Twitch. Given a history of messages, the goal is to determine whether the user suffers from the disorder and the context influencing the mental health problem.

Task 1: Risk Detection of Gambling Disorders — Binary classification.

Task 2: Type of Addiction Detection — Multiclass classification.

3.2. Task 1: Risk Detection of Gambling Disorders

Binary Classification

This task aims to determine whether a user is at high risk (label = 1) or low risk (label = 0) of developing a gambling-related disorder based on their message history. The objective is to enable early detection for timely interventions. Input: A stream of user messages from Telegram and Twitch. Output: A binary label (1 for high risk, 0 for low risk).

3.3. Task 2: Type of Addiction Detection

Multiclass Classification

This task extends Task 1 by requiring the identification of the specific type of gambling addiction for all users, regardless of their risk level (high or low). The available labels are: Betting, Online Gaming, Trading, and Lootboxes.

Input: A stream of user messages from Telegram and Twitch.

Output: One of four addiction types: *Betting*: Gambling on sports or events (e.g., football betting). *Online Gaming*: Games of chance (e.g., roulette, slots). *Trading*: Speculative investments (e.g., cryptocurrencies). *Lootboxes*: Randomized virtual items in video games.

Examples: *Betting (User1)*: "yo de frees no hable"; "pues de pago si hay buenos, pero a mi me ha llevado mi tiempo emplea tu el tuyo". *Online Gaming (User2)*: "24 céntimos con los free spins, Roma no se construyó en un dia". *Trading (User3)*: "Siii fue cuestión de rapidez porqje subio a 60 y rapidio se desinflo". *Lootboxes (User4)*: "y al que le toco sirve para intercambio".

Notes: All users are associated with exactly one addiction type, regardless of risk level. The addiction type prediction evaluated is the one received in the last round. Models must handle noisy, informal, and multilingual text while optimizing for early detection and computational efficiency.

3.4. Dataset

The dataset [15] for the MentalRiskES 2025 shared task is designed for both Task 1 (Risk Detection of Gambling Disorders) and Task 2 (Type of Addiction Detection), with the same dataset used for training and evaluation across both tasks. It consists of user messages collected from Telegram and Twitch, platforms known for informal and community-driven communication. The dataset includes 350 training subjects and 7 trial subjects, totaling 357 subjects, each associated with a stream of messages over time.

Each message entry in the dataset contains the following fields: an id_message (a unique identifier for the message), the message content (the text written by the user), a date (timestamp of the

message in the format “YYYY-MM-DD HH:MM:SS+TZ”), and the platform (either Telegram or Twitch).

An example entry is as follows: ”id_message: 8312066771, message: “Hola buenas alguien sabe si en betfred mandan carta a casa por el registro ?”, date: “2020-09-09 02:50:12+01:00”, platform: “Telegram” ” .

This example illustrates the informal tone and gambling-related content (e.g., reference to “betfred”, a betting platform) typical of the dataset, as well as the multilingual nature, with messages primarily in Spanish.

The dataset is labeled for both tasks. For Task 1, each subject is assigned a binary label indicating their risk level: low risk (label = 0) or high risk (label = 1). The class distribution is shown in Table 1. For Task 2, each subject is also labeled with one of four addiction types: Betting, Online Gaming, Trading, or Lootboxes, regardless of their risk level. The distribution of addiction types is presented in Table 2. Since both tasks share the same dataset, models can leverage the same message streams to jointly or separately address the binary classification of risk (Task 1) and the multiclass classification of addiction types (Task 2), enabling potential synergies in feature extraction and temporal analysis.

Key characteristics of the dataset include its temporal nature, as messages are ordered by timestamp, allowing for streaming or sequential processing to capture evolving user behavior. The informal, noisy, and multilingual nature of the messages (e.g., slang, abbreviations, and platform-specific jargon like “frees” for free spins) poses challenges for natural language processing. Additionally, the dataset’s focus on gambling-related content requires models to identify subtle behavioral cues, such as references to betting platforms, financial risk, or gaming mechanics, to accurately predict risk and addiction types.

Table 1
Class distribution for Task 1.

Class	Train	Trial
Low Risk (0)	178	4
High Risk (1)	172	3
Total	350	7

Table 2
Addiction type distribution for Task 2.

Addiction Type	Train	Trial
Betting	85	2
Online Gaming	104	2
Trading	135	2
Lootboxes	26	1
Total	350	7

4. Methodology

To tackle the challenges of data scarcity and the informal nature of Telegram and Twitch messages in Task 1 of MentalRiskES 2025, we implemented comprehensive preprocessing, data augmentation, and a robust training strategy. Task 1 involves binary classification to identify users as high-risk or low-risk

based on their message sequences. Below, we detail the preprocessing, data processing, and training steps tailored for this task.

4.1. Data Processing

Messages from Telegram and Twitch needed preprocessing to handle informal language, emojis, and decoding errors, followed by task-specific augmentation to improve model performance. We applied these steps for both tasks:

Converting Emojis to Spanish Text Equivalents: Emojis complicated our semantic analysis since they aren't pure text. We converted over 300 emojis to their Spanish text descriptions (e.g., emoji became "smiling face") using Python's `emoji` library. This preserved the meaning while making the data more consistent and easier to analyze.

Manual Correction of Character Errors: During processing, we found incorrect character substitutions, like "o" being replaced with "@@". These issues, stemming from encoding problems or data noise, were manually identified and fixed. Though time-consuming, this step was essential to maintain text integrity before further processing.

Standardizing Repetitive Expressions: Social media data often contains repetitive expressions like "jaja" or "jsjs" (laughter in Spanish). These inconsistent sequences add noise to analysis. We simplified them to standard forms (e.g., both "jaja" and "jsjs" became "ja") using Python processing rules, which reduced complexity and ensured consistency.

Removing Extra Whitespace: Raw texts typically contain unnecessary spaces that increase data size and complicate processing algorithms. We eliminated these using normalization techniques in Python that replace multiple spaces with single ones and remove empty lines, making the data more compact and improving processing efficiency.

Data Augmentation:

> **Task 1 Data Augmentation** For Task 1's balanced dataset (178 low-risk, 172 high-risk in training), we linked labels with message sequences from the ground truth file. For low-risk subjects, we randomly selected (n) messages (from 1 to total count) in each epoch, using only the first (n) to encourage early negative predictions. For high-risk subjects, with 0.3 probability, we added up to 10 low-risk messages at the beginning to prevent overfitting. A PyTorch dataset class dynamically created augmented samples, handling variable-length sequences.
>
> **Task 2 Data Augmentation** For Task 2's imbalanced dataset, especially Lootboxes (only 26 training samples), we translated messages (Spanish to English to French to Spanish) to create paraphrased versions, doubling our sample count. We created new user entries (e.g., adding "augment" to IDs) and saved the augmented data to a JSON file. We checked cached data to avoid redundant processing.

This unified approach ensured quality input and increased training sample diversity, addressing data scarcity and imbalance while maintaining temporal and semantic integrity.

4.2. Task 1: Risk Detection Solution

Our solution for Task 1 employs a bidirectional Long Short-Term Memory (LSTM) classifier with an attention mechanism to process 768-dimensional message embeddings generated by the RoBERTuito model (`pysentimiento/robertuito-sentiment-analysis`), fine-tuned on the TASS 2020 corpus for Spanish sentiment analysis. Messages are processed in rounds, and the first LSTM output equaling 1

(high-risk) at any timestep classifies the user as addicted, enabling early risk detection in informal, noisy Telegram and Twitch message sequences. This approach, combined with Group Distributionally Robust Optimization (GroupDRO), soft labels, and targeted data augmentation, addresses class imbalance and ensures robust performance. The use of Bi-LSTM for sequential text analysis in mental health prediction has been shown to be effective in prior work [16]. The key technical strengths are:

1. **Robust Temporal Modeling with Round-Based Classification**: The bidirectional LSTM processes sequences of RoBERTuito embeddings, capturing temporal dependencies in both forward and backward directions. Each message is encoded into a 768-dimensional embedding, and the LSTM generates outputs for each timestep (round) by concatenating forward and backward hidden states. An attention mechanism (a linear layer mapping 2×h_size to 1) computes softmax weights to emphasize relevant timesteps, followed by a classifier layer (2×h_size to 2) with sigmoid activation to produce probabilities. During inference, predictions are made for all timesteps, and the first timestep where the output equals 1 (high-risk) classifies the user as addicted; otherwise, the user is labeled low-risk (0). This round-based approach, leveraging the first high-risk prediction, ensures early detection of addiction signals, critical for real-time monitoring. The RoBERTuito model's pre-training on Spanish informal text enhances feature extraction, enabling robust handling of varied sequence lengths and informal language.

2. **Enhanced Generalization with Soft Labels**: To address ambiguity in informal messages, we use soft labels: [0.95, 0.05] for low-risk and [0.05, 0.95] for high-risk. This mitigates the impact of messages with mixed risk signals, encouraging smoother decision boundaries. The cross-entropy loss, computed with these probabilistic targets, enables the model to learn nuanced patterns, reducing overfitting to noisy annotations and improving generalization across diverse user behaviors in the round-based setting.

3. **Addressing Class Imbalance with Augmentation and GroupDRO**: Despite the near-balanced dataset (178 low-risk, 172 high-risk), subtle imbalances in risk signal distribution require robust handling. We augment high-risk samples (label 1) by prepending up to 10 low-risk messages with a 70% probability per epoch (threshold 0.6), implemented via a custom PyTorch dataset ('EmbDatasetRNNAug'). This increases training diversity, simulating scenarios where high-risk signals appear later in rounds, preventing overfitting to early cues. For low-risk samples, we randomly select a subsequence of n messages (1 to total messages) per epoch, promoting early negative predictions. GroupDRO complements this by dynamically adjusting class weights based on loss:

$$q'_g = q_g^{(t-1)} \exp\left(\eta \mathcal{L}\left(f_\theta, S_g\right)\right), \quad q_g^{(t)} = \frac{q'_g}{\sum_{g' \in \mathcal{G}} q'_{g'}},$$

$$\mathcal{L}_{GDRO}\left(f_\theta, (X, Y)\right) = \sum_{g \in \mathcal{G}} q_g^{(t)} \mathcal{L}\left(f_\theta, S_g\right),$$

where $q_g^{(t)}$ is the weight for group g (high-risk or low-risk), initialized uniformly ($q_g^{(0)} = 1/|\mathcal{G}|$), and $\eta = 0.1$. The loss $\mathcal{L}\left(f_\theta, S_g\right)$ prioritizes harder-to-learn samples, ensuring balanced optimization. This dual strategy enhances performance for high-risk users, critical for accurate addiction detection. The Adam optimizer is used to update model parameters.

4. **Early Risk Detection via Round-Based Output and Augmentation**: The round-based classification, where the first LSTM output equaling 1 triggers a high-risk label, is optimized for early addiction detection. The augmentation strategy—truncating low-risk sequences and enriching high-risk sequences with low-risk messages—trains the model to identify high-risk signals in early rounds. Specifically, for Task 1's balanced dataset (178 low-risk, 172 high-risk in training), labels were associated with user message sequences from the ground truth file. For low-risk subjects, we randomly selected n messages (ranging from 1 to the total number of messages) at each epoch, using only the

first n messages to encourage early negative predictions. For high-risk subjects, with a 30% probability, we prepended up to 10 low-risk messages to the sequence to prevent overfitting to early high-risk signals. A custom PyTorch dataset class dynamically generated these augmented samples, maintaining efficiency and enabling generalization across sporadic risk signals. The bidirectional LSTM's global context, combined with the attention mechanism, ensures that early predictions are informed by the entire sequence, enhancing accuracy.

The dataset (357 training subjects) is preprocessed as described in Section 4.1 and split into 539 samples (364 low-risk, 175 high-risk) by dividing low-risk sequences into two subsequences. A custom PyTorch dataset EmbDatasetRNNAug dynamically augments data: low-risk sequences are truncated to a random length (1 to total messages) per epoch, and high-risk sequences are prepended with up to 10 low-risk messages with 70% probability (threshold=0.6). We split the data into 80% training and 20% validation using a fixed seed (2112). RoBERTuito generates 768-dimensional CLS embeddings (max length=96 tokens) on a GPU in evaluation mode.

The model is trained for 150 epochs using the Adam optimizer (learning rate=1e-4) with batch sizes of 2, 4, or 8. Soft labels ([0.95, 0.05] for low-risk, [0.05, 0.95] for high-risk) are used with cross-entropy loss to handle ambiguity. GroupDRO adjusts class weights dynamically ($\eta = 0.1$) to prioritize harder samples. We test hidden sizes of 32, 64, and 96, with the best validation Macro F1 (0.91) achieved at hidden size=32, batch size=2, epoch 117.

Figure 1 presents the training and validation Macro F1 scores over 120 epochs, reflecting the model performance for Run 0 with data augmentation. The plot shows a rapid increase in Train Macro F1 (blue line) to approximately 0.95 within the first 20 epochs, stabilizing thereafter, while the Val Macro F1 (orange line) rises to around 0.91 with noticeable fluctuations, indicating potential overfitting or variability in validation data. This suggests that the augmentation strategy effectively boosts training performance, though validation stability could benefit from further regularization.

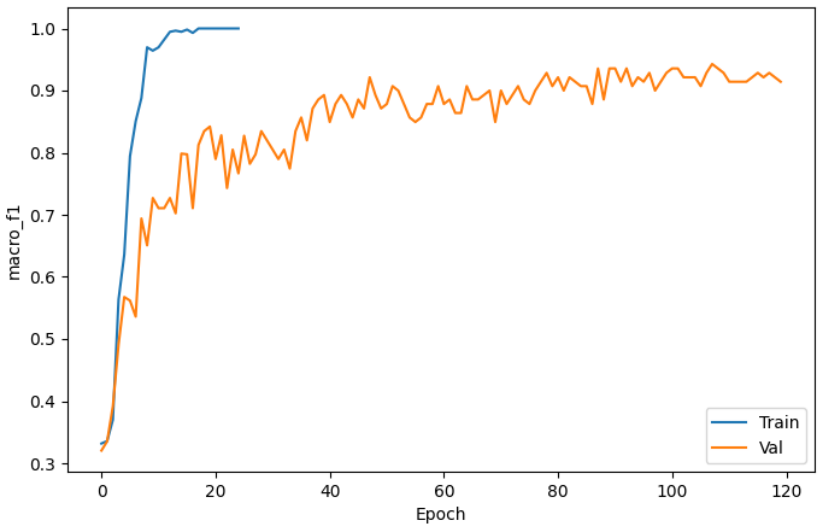

Figure 1: Training and Validation Macro F1 Scores for Task 1

The training pipeline for Task 1, shown in Figure 2, illustrates the sequential flow from raw message preprocessing to round-based prediction.

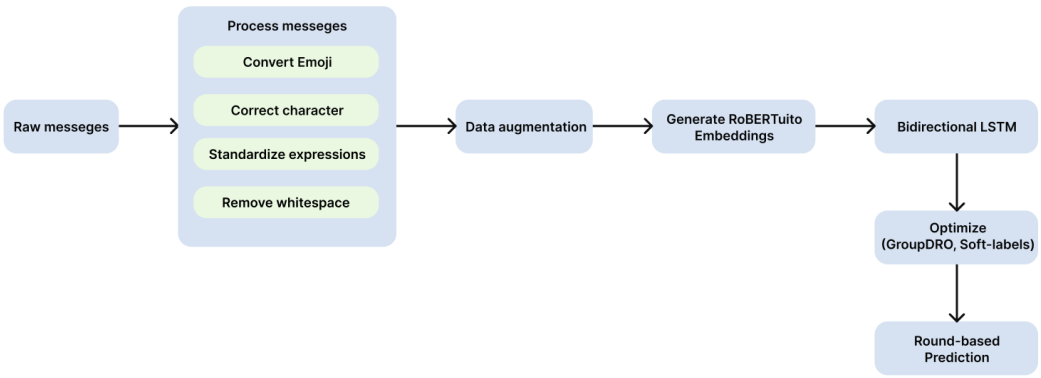

Figure 2: Training pipeline for Task 1: Risk Detection Solution.

Table 3
Hyperparameters for Task 1. Run 0 includes data augmentation, while Runs 1 and 2 do not.

Run	Batch Size	Hidden Size	Val Macro F1
0 (with augmentation)	2	32	0.91
1 (no augmentation)	2	64	0.7612
2 (no augmentation)	4	64	0.7218

This solution excels in handling informal, noisy data, achieving high Macro F1 scores and balanced performance across classes, particularly for early risk detection in high-risk users, thanks to targeted augmentation and robust optimization. The use of data augmentation in Run 0 notably improves performance (Macro F1 of 0.91) compared to Runs 1 and 2 (0.7612 and 0.7218, respectively), highlighting the effectiveness of the augmentation strategy in enhancing early risk detection.

4.3. Task 2: Addiction Type Detection Solution

Our solution for Task 2 fine-tunes the roberta-base-bne model (`PlanTL-GOB-ES/roberta-base-bne`) for multiclass classification to identify addiction types (Betting, Online Gaming, Trading, Lootboxes), addressing severe class imbalance, particularly for Lootboxes (26 training samples), through back-translation augmentation and class-weighted loss. The model processes concatenated message sequences, leveraging robust feature extraction and optimization strategies to achieve high performance across all classes. The key technical strengths are:

1. **Robust Feature Extraction**: The RoBERTa model, pre-trained on Spanish corpora, excels at extracting nuanced linguistic features from concatenated message sequences. Messages per subject are tokenized and padded to a maximum length of 256 tokens, with special tokens and attention masks to handle variable-length inputs. The model's transformer architecture captures contextual relationships in informal Telegram and Twitch messages, enabling accurate differentiation of addiction types despite linguistic variability.

2. **Effective Handling of Class Imbalance**: To address the severe imbalance (Lootboxes: 26 samples), we double the Lootboxes samples via back-translation (Spanish to English to French to Spanish) using MarianMT models (`Helsinki-NLP/opus-mt-es-en`, `en-fr`, `fr-es`). This generates paraphrased messages, preserving semantic content while increasing training diversity. New user entries are created with an "augment" suffix, and augmented data is cached in a JSON file for efficiency. During training, a class-weighted cross-entropy loss prioritizes the minority class with weights [1.0, 1.0, 1.0, 1.5] for Betting, Online Gaming, Trading, and Lootboxes, respectively, ensuring robust performance

on underrepresented samples.

3. **Optimized Hyperparameter Tuning**: Hyperparameter optimization with Optuna over 20 trials searches learning rates (3e-4 to 5e-4), batch sizes (2, 4), epochs (4 to 10), and weight decay (0.05 to 0.1), maximizing the average Macro F1 score across 5-fold stratified cross-validation. The best configuration (learning rate 3.1e-4, batch size 4, 7 epochs, weight decay 0.06) is selected based on cross-validation performance, with early stopping based on Macro F1 to prevent overfitting, ensuring optimal model convergence.

4. **Generalization Across Classes**: The model employs 5-fold stratified cross-validation to maintain class distribution across splits, ensuring robust evaluation. Mixed precision training (FP16) on GPU enhances efficiency, while gradient accumulation (2 steps) stabilizes training with small batch sizes. Dropout probabilities (0.2 for hidden and attention layers) mitigate overfitting, promoting generalization across diverse addiction types, particularly for the minority Lootboxes class.

The model was trained using 5-fold stratified cross-validation with a fixed random seed (42), employing a custom `WeightedTrainer` to apply class-weighted loss. The dataset was tokenized with a maximum length of 256, and training leveraged 500 warmup steps and gradient accumulation, with the best model saved based on Macro F1. The hyperparameters corresponding to this best model are listed in Table 4

The training pipeline for Task 2, shown in Figure 3, illustrates the process from raw message preprocessing to addiction type prediction.
This solution effectively handles class imbalance and informal language, delivering robust performance across all addiction types, particularly for the minority Lootboxes class, due to targeted augmentation and optimized training.

Table 4
Best hyperparameters for Task 2.

Hyperparameter	Value
Learning Rate	3.1e-4
Batch Size	4
Epochs	7
Weight Decay	0.06

5. Evaluation

The tasks are evaluated in a round-based setup, with one message per subject per round submitted via HTTPS POST requests through a provided API. For Task 1, the first positive prediction (high risk) is considered for evaluation. For Task 2, the addiction type predicted in the last round is evaluated. Performance is assessed using a combination of classification metrics, latency metrics, and computational efficiency metrics.
Classification Metrics:

Macro F1: The macro-averaged F1 score calculates the harmonic mean of precision and recall across all classes (high/low risk for Task 1, and Betting/Online Gaming/Trading/Lootboxes for Task 2), treating each class equally regardless of its frequency. This metric is particularly useful for imbalanced datasets, ensuring that performance on minority classes (e.g., Lootboxes) is not overshadowed by majority classes (e.g., Trading).

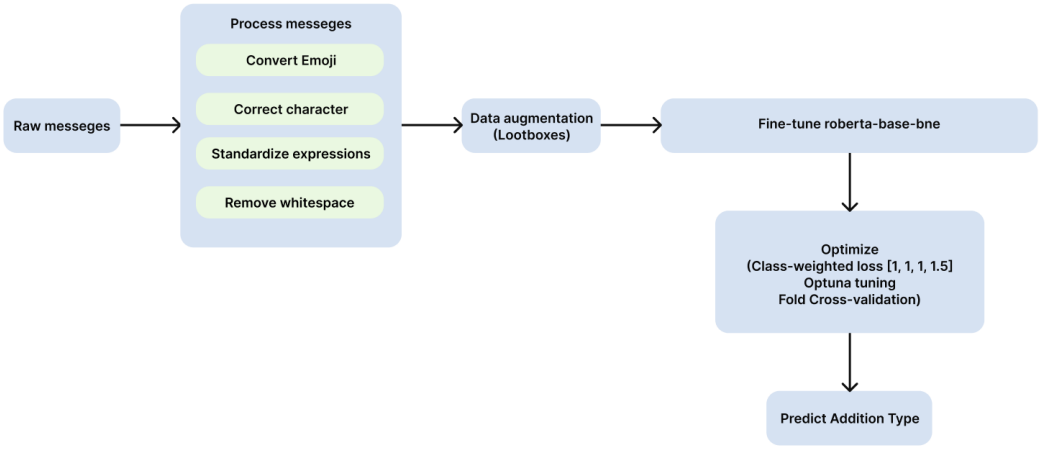

Figure 3: Training pipeline for Task 2: Addiction Type Detection Solution.

Macro F1_c: An enhanced version of Macro F1, this metric adjusts for cost-sensitivity by incorporating a cost matrix that penalizes false negatives (e.g., missing a high-risk user) more heavily than false positives, reflecting the real-world importance of early detection in mental health applications.

Latency Metrics:

ERDE$_5$: The Early Risk Detection Error at 5 messages measures the error in predicting the correct class (risk level or addiction type) within the first 5 messages per subject. It penalizes late or incorrect predictions with an error score that increases with delay, emphasizing the need for early and accurate detection to enable timely interventions.

ERDE$_{30}$: Similar to ERDE$_5$, this metric evaluates the error over the first 30 messages per subject. It provides a broader window for assessment, allowing models more data to refine predictions, but still prioritizes early detection by penalizing delays beyond 30 messages.

Computational Efficiency Metrics:

RAM: The memory usage of the model during training and inference, measured in gigabytes, reflects the scalability of the solution for deployment on resource-constrained environments.

CPU: The processor utilization, measured as a percentage or in computational units, indicates the computational demand of the model, critical for real-time processing of message streams.

Carbon Emissions: Quantified using the CodeCarbon tool, this metric estimates the environmental impact in kilograms of CO_2 equivalent, encouraging the development of energy-efficient models aligned with sustainable AI practices.

6. Results and Discussion

6.1. Task 1

Our best submission (Run 0) for Task 1, leveraging a bidirectional LSTM with an attention mechanism and RoBERTuito embeddings, achieved a Macro F1 score of 0.403, placing 21st among all participants (Table 5). While this performance reflects robust modeling of temporal dependencies and informal

language, the score is notably lower than our validation Macro F1 of 0.91, suggesting overfitting or a distribution shift between training and test datasets. Higher ERDE5 values, particularly in Run 1 (0.577), indicate delayed positive predictions, likely due to the model's sensitivity to early risk signals being insufficiently optimized. Preprocessing steps, such as replacing emojis with Spanish textual equivalents and standardizing repetitive substrings, enhanced dataset quality, while augmentation strategies—like prepending low-risk messages to high-risk samples and truncating low-risk sequences—improved robustness to varied sequence lengths. However, GroupDRO optimization and hyperparameter tuning did not fully address challenges in early detection, particularly for high-risk users. Future improvements could focus on mitigating overfitting through stronger regularization, diversifying augmentation to better simulate test conditions, and fine-tuning the attention mechanism to prioritize early risk signals, thereby reducing the gap between validation and test performance.

6.2. Task 2

Our best submission (Run 2) for Task 2, fine-tuned with the `roberta-base-bne` model and back-translation augmentation for the Lootboxes class, achieved a Macro $F1_c$ (Macro F1 with class-weighted corrections) of 0.822, ranking 10th among participating teams (Table 6). While this represents a competitive performance, there remains room for improvement. The relatively high Macro $F1_c$ reflects our system's ability to detect different addiction classes, supported by the class-weighted cross-entropy loss and 5-fold cross-validation. However, precision variations across runs (e.g., Run 1 with ERDE5 0.577) indicate minor false positives, likely due to the back-translation strategy introducing subtle semantic shifts. Hyperparameter optimization with Optuna (learning rate 3.1e-4, batch size 4, 7 epochs) effectively balanced performance across addiction types.

Table 5
Risk detection results.

Rank	Team	Run	Accuracy	Macro F1	ERDE5	ERDE30
1	UNSL	2	0.569	0.567	0.639	0.389
3	I2C-UHU-Rigel	1	0.556	0.551	0.600	0.284
21	PuxAI	0	0.538	0.403	0.434	0.302
22	PuxAI	2	0.525	0.396	0.377	0.283
25	PuxAI	1	0.525	0.367	0.577	0.262

Table 6
Addiction type detection results.

Rank	Team	Run	Accuracy	Macro $F1_c$	ERDE5	ERDE30
1	PLN_PPM_ISB	1	0.569	0.927	0.412	0.248
3	ELiRF-UPV	1	0.550	0.887	0.579	0.402
10	PuxAI	2	0.531	0.822	0.370	0.277
19	PuxAI	0	0.531	0.722	0.435	0.299
23	PuxAI	1	0.525	0.713	0.577	0.262

6.3. Practical Considerations

The transformer-based models employed, particularly the LSTM for Task 1 and RoBERTa for Task 2, are computationally intensive. The preprocessing steps (emoji mapping, decoding correction, text normalization) and augmentation strategies (negative sample splitting for Task 1, back-translation for Task 2) enhanced model input quality but increased processing demands. Applying distillation could reduce resource usage, especially for deployment. Classifier calibration, such as Platt scaling, would improve confidence scores, addressing the delayed predictions indicated by ERDE5 metrics and

supporting real-world application.

The performance variations across Task 1 runs stem primarily from the use of data augmentation. Run 0, which employs dynamic augmentation via `EmbDatasetRNNAug` (truncating low-risk sequences and prepending up to 10 low-risk messages to high-risk sequences with 70% probability), mitigates class imbalance and enhances prediction balance across low-risk and high-risk classes. This results in the highest validation Macro F1 (0.91, Table 3) compared to Runs 1 and 2 (0.7612 and 0.7218, respectively), which lack augmentation and thus exhibit poorer generalization on imbalanced data.

In Task 2, Run 2 achieves the highest Macro $F1_c$ (0.822, Table 6) due to targeted data augmentation for the underrepresented Lootboxes class (26 samples) and a slight increase in its class weight. Specifically, back-translation (Spanish to English to French to Spanish) doubles Lootboxes samples, enhancing training diversity, while the class-weighted cross-entropy loss assigns a higher weight (1.5 vs. 1.0 for other classes) to prioritize Lootboxes. In contrast, Runs 0 and 1 (Macro $F1_c$ of 0.722 and 0.713, respectively) lack these adjustments, leading to rapid loss convergence but suboptimal performance, even when selecting epochs with moderate loss, due to insufficient handling of class imbalance.

6.4. Environmental Impact

The carbon footprint of submission is quantified using CodeCarbon, estimating CO_2-equivalent (CO_2e) emissions on an NVIDIA RTX 4060 GPU. For Task 1, mean emissions are 0.0000675 kg CO_2e per run (standard deviation: 0.0000994 kg CO_2e), reflecting efficient GPU utilization. For Task 2, mean emissions are slightly higher at 0.0000741 kg CO_2e (standard deviation: 0.0001000 kg CO_2e), due to the computational complexity of multiclass classification. These low emissions align with sustainable AI practices.
For training, emissions are higher due to iterative computations across multiple epochs. Task 1's bidirectional LSTM training (150 epochs, data augmentation) and Task 2's fine-tuned roberta-base-bne (4–7 epochs, back-translation, Optuna tuning) on the RTX 4060 GPU likely produce emissions several times the submission values. Task 2's higher complexity suggests greater training emissions than Task 1.Despite these relatively modest environmental impacts, model pruning or optimized hyperparameter search could further reduce impact. Further reductions could be achieved through: using early stopping or Bayesian optimization instead of exhaustive Optuna trials to minimize redundant computations or leveraging lower precision formats (e.g., INT8) during inference to decrease GPU energy consumption.

7. Error Analysis

To understand the limitations of our proposed solutions for Task 1 (Risk Detection) and Task 2 (Addiction Type Detection), we conducted an in-depth analysis of failed predictions, focusing on their characteristics, potential causes, and lessons learned. This section highlights key issues contributing to suboptimal performance, proposes improvements, and discusses challenges specific to the MentalRiskES 2025 dataset.

7.1. Characterization of Failed Predictions

For Task 1, the bidirectional LSTM with RoBERTuito embeddings achieved a test Macro F1 of 0.403, significantly lower than the validation Macro F1 of 0.91. Analysis of misclassified cases revealed that the model struggled with long message sequences, particularly for high-risk users whose risk signals appeared later in the sequence. The high ERDE5 scores (e.g., 0.577 in Run 1) indicate delayed positive predictions, suggesting that the model failed to detect early risk patterns in some subjects. For Task 2, the roberta-base-bne model achieved a Macro $F1_c$ of 0.822, but precision varied across runs, with false positives observed for the minority Lootboxes class. These errors were often linked to semantic

ambiguities in informal messages, where paraphrased augmented data introduced subtle shifts in meaning, confusing the model's ability to distinguish addiction types.

7.2. Key Issues and Causes

Small Hidden Size in LSTM (Task 1): The bidirectional LSTM used a hidden size of 64 (Table 3), which likely limited its capacity to model complex temporal dependencies in long message sequences. This constrained representational power may have hindered the model's ability to capture nuanced risk signals, especially in extended sequences where early cues were sparse.

Overfitting Due to Limited Training Data: Despite data augmentation (e.g., negative sample splitting for Task 1 and back-translation for Task 2), the training dataset (350 subjects) remained relatively small. This led to overfitting, as evidenced by the significant gap between validation (0.91 Macro F1) and test (0.403 Macro F1) performance for Task 1. The augmentation strategies increased sample diversity but were insufficient to fully generalize to the test distribution.

Data Quality and Platform-Specific Challenges: Messages from Telegram and Twitch exhibited high levels of noise, including slang, abbreviations, and platform-specific jargon (e.g., "frees" for free spins). These characteristics diverged from the cleaner, more standardized Spanish corpora used to pretrain RoBERTuito and roberta-base-bne, posing challenges for semantic understanding. Additionally, inconsistencies in message length and temporal distribution further complicated early risk detection.

Imbalance and Augmentation Artifacts (Task 2): While back-translation doubled the Lootboxes samples, it occasionally introduced paraphrasing errors, leading to false positives. For example, translated messages sometimes lost gambling-specific context, reducing their discriminative power for addiction type classification.

8. Conclusions

PUXai's solutions for the MentalRiskES 2025 shared task effectively addressed the challenges of early gambling disorder detection using Spanish language-specific transformer models paired with advanced data augmentation strategies. For Task 1, the combination of RoBERTuito embeddings with a bidirectional LSTM and GroupDRO optimization demonstrated robust handling of temporal dependencies in noisy Telegram and Twitch messages. For Task 2, fine-tuning roberta-base-bne with back-translation and Optuna-based hyperparameter tuning successfully mitigated class imbalance, enabling accurate identification of addiction types. Preprocessing techniques, including emoji mapping and text normalization, were critical in enhancing input quality, while augmentation strategies ensured resilience against data scarcity and informal language. Although our approaches did not secure top rankings, they highlight the potential of tailored language models and strategic data processing for mental health risk detection in real-world, noisy environments.

Generative AI Declaration

The authors declare that all ideas, analyses, and research contributions presented in this paper are their own. Generative AI tools (e.g., ChatGPT) were used solely to assist in drafting and refining the text. All AI-assisted content has been thoroughly reviewed, critically evaluated, and verified by the authors, who take full responsibility for the final manuscript.s

References

[1] M. Sparrow, R. Volberg, J. Rehm, Is gambling becoming a public health crisis?, Harvard Magazine (2025). URL: https://www.harvardmagazine.com/2025/03/harvard-research-gambling-public-health-crisis.

[2] E. Smith, J. Peters, N. Reiter, Automatic detection of problem-gambling signs from online texts using large language models, PLOS Digital Health 3 (2024) e0000605. URL: https://journals.plos.org/digitalhealth/article/10.1371/journal.pdig.0000605. doi:10.1371/journal.pdig.0000605.

[3] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2019: Early risk prediction on the internet, in: CEUR Workshop Proceedings, volume 2380, 2019, p. 248. URL: https://ceur-ws.org/Vol-2380/paper_248.pdf.

[4] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, 2022. URL: https://arxiv.org/abs/2111.09453. arXiv:2111.09453.

[5] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[6] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156. doi:10.26342/2022-68-3.

[7] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.-T. Lu, C. C. Aggarwal, J. Pei, Y. Zhou, A comprehensive survey on data augmentation, 2025. URL: https://arxiv.org/abs/2405.09591. arXiv:2405.09591.

[8] Q. Xu, Y. Hong, J. Chen, J. Yao, G. Zhou, Data augmentation via back-translation for aspect term extraction, in: 2023 International Joint Conference on Neural Networks (IJCNN), 2023, pp. 1–8. doi:10.1109/IJCNN54540.2023.10191183.

[9] S. Sagawa, P. W. Koh, T. B. Hashimoto, P. Liang, Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL: https://arxiv.org/abs/1911.08731. arXiv:1911.08731.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.

[11] A. Parfenova, M. Clausel, Risk prediction of pathological gambling on social media, 2024. URL: https://arxiv.org/abs/2403.19358. arXiv:2403.19358.

[12] A. Chouhan, A. Halgekar, A. Rao, D. Khankhoje, M. Narvekar, Sentiment analysis of twitch.tv livestream messages using machine learning methods, in: 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2021, pp. 1–5. doi:10.1109/ICECCT52121.2021.9616932.

[13] A. M. Mármol-Romero, P. Álvarez-Ojeda, A. Moreno-Muñoz, F. M. P. 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.

[16] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, 2015. URL: https://arxiv.org/abs/1508.01991. arXiv:1508.01991.