VerbaNexAI at MentalRiskES 2025: Early Detection of Gambling Disorders using Transformer Architectures and Machine Learning Models

Jeison D. Jimenez[1,*], Jairo E. Serrano[1], Juan C. Martinez-Santos[1] and Edwin Puertas[1]

[1]*Universidad Tecnologica de Bolivar, School of Digital Transformation, Cartagena de Indias 130010, Colombia.*

Abstract

Gambling disorder represents a significant public health challenge with severe psychological and socioeconomic consequences, affecting approximately 80 million individuals worldwide. Early detection and intervention are crucial for mitigating its harmful effects. In this paper, we present the approach of VerbaNexAI to the MentalRiskES 2025 shared task, focusing on the early detection of gambling disorders in Spanish social media content. Our methodology leverages transformer-based embeddings combined with traditional machine learning algorithms, applied to a novel dataset of Spanish-language user profiles from Telegram and Twitch platforms. We implemented a comprehensive pipeline including text preprocessing, contextual embedding generation using Spanish BERT models, class balancing through random oversampling, and systematic model selection. Our LightGBM classifier demonstrated moderate classification performance for risk detection but excelled in early detection capabilities among all participating teams. Our Logistic Regression model delivered robust results across multiple evaluation criteria for addiction type classification. This research contributes to the field by demonstrating the effectiveness of combining transformer architectures with traditional machine learning for early detection of gambling disorders in Spanish text, with particular emphasis on timeliness of detection as a critical factor for effective mental health monitoring and intervention.

Keywords

Mental Risk, Gambling, Embedding, Transformers, Machine Learning

1. Introduction

Mental health represents one of the most pressing healthcare challenges worldwide: one in eight people suffers from a mental disorder, yet the majority lack access to adequate treatment [1]. The health crisis caused by COVID-19 intensified this issue, increasing the global prevalence of anxiety and depression by 25% during its first year and thereby intensifying preexisting structural deficiencies in care services [2]. Suicide remains the fourth leading cause of death among individuals aged 15 to 29, accounting for over 700,000 deaths annually [3].

In this context of collective psychological vulnerability, gambling disorder has emerged as a significant public health threat: recent Lancet Commission research estimates that approximately 80 million individuals meet diagnostic criteria for this disorder. At the same time, up to 450 million engage in harmful gambling behaviours worldwide [4]. Conceptualized in the DSM-5 as persistent, recurrent problematic gambling behaviour leading to clinically significant impairment or distress [5], gambling disorder exhibits the highest suicide rate among addictive disorders, with clinically relevant rates of suicidal ideation (31.6 %) and suicide attempts (13.2 %) [6]. Its socioeconomic impact is equally alarming: social-cost analyses estimate per-adult harms ranging from USD 16 to USD 36,144 annually, encompassing economic, relational, and psychological consequences [7]. Paradoxically, only one in five affected individuals seeks professional assistance [8].

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

jalvear@utb.edu.co (J. D. Jimenez); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0009-0001-0134-8426 (J. D. Jimenez); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

Implementing automated systems for mental-health risk detection on digital platforms constitutes a scalable methodology for early intervention. The CLEF eRisk lab has driven significant advances in early risk prediction in English [9]; however, Spanish remains under-represented. To address this gap, IberLEF [10] incorporated the Early-Risk Identification task (MentalRiskES) into SEPLN (the Spanish Society for Natural Language Processing) in 2023 [11] and 2024 [12], thereby establishing a methodological framework for mental-risk evaluation in Spanish. In this third edition [13], two subtasks have been defined:

- Task 1: Risk Detection of Gambling Disorders. This online binary classification task requires systems to decide, for each incoming message in a chronological stream of Telegram and Twitch user comments, whether the user is at high risk (label = 1) or low risk (label = 0) of developing a gambling-related disorder. Early detection is paramount: we evaluated performance on classification accuracy and the promptness with which we issued a correct "high risk" label once sufficient evidence was available.
- Task 2: Type of Addiction Detection. This risk-conditioned multiclass classification task requires systems to assign each user exactly one addiction type: Betting, Online Gaming, Trading, or Lootboxes, based on their message history, regardless of prior risk flagging. The final prediction submitted in the last round is used for evaluation, emphasizing both label correctness and decision timeliness.

2. Related Work

In recent years, the automatic detection of pathological gambling using machine learning and natural language processing methods has gained significant attention. Researchers have explored various approaches to identify gambling disorder signals in text data, leveraging linguistic patterns and behavioral indicators. This section provides an overview of key studies in this domain, highlighting the most effective approaches and results obtained.

ELiRF-UPV [14] conducted one of the most successful studies in pathological gambling detection, employing Support Vector Machines (SVM) with TF-IDF features for the eRisk 2023 shared task. Their approach prioritized handling long texts effectively, achieving remarkable results with an F1 score of 0.935 on validation data and perfect precision (1.000) in test scenarios. Their system ranked first among 49 submissions in the eRisk 2023 challenge, detecting 75% of gamblers within the first 10 posts, demonstrating the continued viability of classical machine learning methods even as transformer models gain popularity.

NLP-UNED-2 [15] introduced an innovative approach for the eRisk 2022 challenge using Approximate Nearest Neighbors (ANN) for dataset relabeling to convert user-level annotations to message-level labels. They refined the training data using Universal Sentence Encoder embeddings and HNSW graphs. They implemented RNN-based models that achieved impressive results, ranking second in the eRisk 2023 decision-based evaluation with high precision (0.896) and recall (0.922). Their latency-weighted F1 score of 0.877 demonstrated the effectiveness of neural networks with properly labeled data.

SINAI [16] presented an approach to the eRisk 2023 challenge, leveraging pre-trained Transformer-based models (RoBERTa-Large and XLM-RoBERTa-Large) combined with Long Short-Term Memory (LSTM) architectures to analyze social media posts in sequential order. Their methodology emphasized comprehensive data preprocessing, including normalization of text (replacing URLs, emojis, and special characters), handling imbalanced datasets through sub-sampling, and integrating sequential modeling for early detection. The team implemented five systems: four utilizing Transformer-based models with feedforward neural networks (FFNN) and a fifth introducing an innovative hybrid architecture combining RoBERTa with LSTM to capture temporal dependencies in user posts. Despite ranking 7th out of 49 submissions with an F1 score of 0.126, SINAI achieved notable performance in recall and early detection metrics (ERDE50 = 0.020–0.029), demonstrating strong capability in identifying high-risk users in early stages. Their work highlights the potential of combining transformer architectures with sequential models for temporal analysis of user behavior.

UNSL [17] leveraged transformer-based models for early gambling detection in the eRisk 2023 challenge, employing BERT architectures with domain-specific vocabulary enrichment from external models. Their approach incorporated a decision policy based on historical predictions, which optimized early classification by applying thresholds and delay parameters. Their models achieved competitive results in the challenge, particularly excelling in decision-based metrics (F1, ERDE50) and runtime efficiency, with UNSL ranking among the fastest teams in the competition.

UNED-NLP [18] applied an Approximate Nearest Neighbors method with semantic embeddings for the eRisk 2022 challenge, achieving competitive results in early detection (ERDE50 = 0.018). Their methodology involved constructing a reference database of labeled user profiles and classifying new users based on their nearest neighbors in the embedded space. This lightweight approach demonstrated scalability and low runtime latency, suggesting its utility in resource-constrained environments.

Beyond academic competitions, researchers [19] have developed predictive models using player account data, such as transaction logs and betting patterns. Their model achieved strong discriminatory power (AUC > 0.85) in identifying at-risk gamblers by analyzing features like loss-to-win ratios, betting frequency spikes, and late-night gambling activity. This approach complements text-based methods by incorporating behavioral data that may reveal gambling disorders before they manifest in social media discourse.

Research on pathological gambling detection highlights diverse methodologies with complementary strengths. Classical machine learning models (e.g., SVM) demonstrate robust performance when rigorously implemented. At the same time, neural architectures with engineered features enable effective early identification. Hybrid frameworks, such as SINAI's combination of transformers and LSTMs, integrate semantic analysis with temporal modeling, underscoring the value of combined approaches for precise and timely detection.

3. Data

For the MentalRiskES challenge, authors compiled a novel dataset, comprising 517 anonymized Spanish-language user profiles (7 trial, 350 training, 160 testing)[20]. The data were sourced from public Telegram groups dedicated to gambling-related topics, where users exchange messages about wagering activities, and from live-chat streams on Twitch discussing gambling. The organisers then extracted, anonymised, annotated, and curated each user's message history to support two subtasks. As previously outlined, they divided the corpus into three sets for trial, training, and test. Table 1 provides a synopsis of the distribution of users by risk category for Task 1. In contrast, Table 2 offers an overview of the distribution by gambling-type label for Task 2.

Table 1
Distribution of Users by Risk Level for Task 1 (Gambling Disorder Detection)

Risk Level	Users
High Risk	172
Low Risk	178

Table 2
Distribution of Users by Type of Gambling for Task 2 (Addiction Type Detection)

Type of Gambling	Users
Trading	135
Online gaming	104
Betting	85
Lootboxes	26

4. Architecture

In this section, we describe the architecture of our system for the early detection of gambling disorders based on user comments from Telegram and Twitch. As Figure 1 shows, the processing workflow consists of four main stages: preprocessing, where we cleaned messages, normalized, and standardized; feature extraction, in which we generated contextual embeddings through a Spanish Transformer model and relevant lexical features are incorporated; training and validation, which involves comparing various Machine Learning algorithms through cross-validation to ensure robustness; and model evaluation and selection, based on metrics including Accuracy, Recall, F1, Precision, Kappa, and MCC, as well as the earliness of detection.

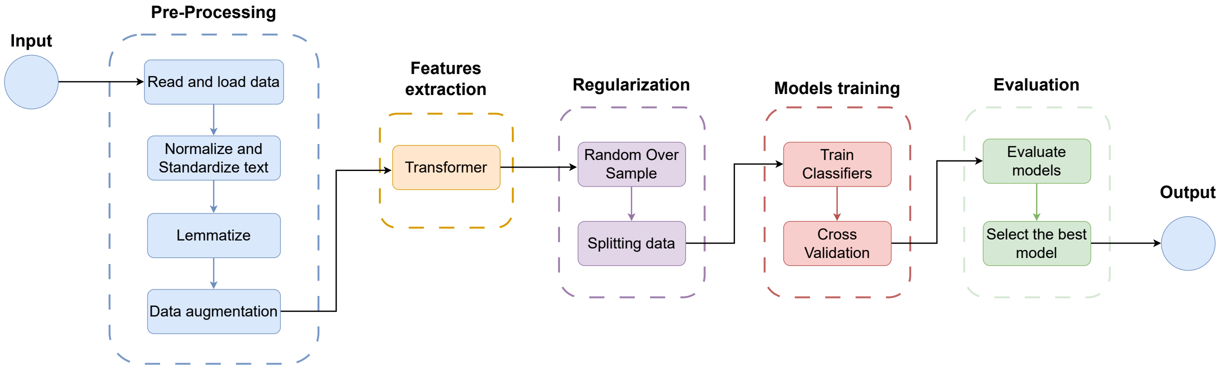

Figure 1: System Pipeline [21].

4.1. Pre-Processing

The first step in our pre-processing pipeline is to convert all text to lowercase. It standardizes the text and ensures consistency in our analysis. Next, we replace email addresses and usernames with tags [EMAIL] and [USER]. Then, we replace emojis with textual descriptions and remove extra spaces. We also normalize quotation marks and reduce repeated punctuation to a maximum of three consecutive instances.

Subsequently, we tokenized the text to segment it into lexical units that facilitate the application of lemmatization on the comments, thus reducing words to their base forms. We then concatenate the messages for each user, using the [SEP] separator to distinguish between different interactions and maintain the conversational structure. This process is implemented for both Task 1 and Task 2, as they share the same dataset of messages.

Furthermore, for Task 2, we implement data augmentation techniques specifically in the Lootboxes class because this category presents a marked imbalance compared to the others, having only 26 instances, as shown in Table 2. To mitigate this limitation, we generated 59 additional message instances using three techniques: back translation (Spanish-English-Spanish), synonym substitution via Spanish WordNet, and paraphrasing by combining both methods. The augmentation was applied at the message level rather than creating synthetic user profiles, increasing the Lootboxes class from 26 to 85 instances. This approach allowed us to expand the training set while preserving the semantic integrity of the original examples and achieving better class balance across all gambling categories.

4.2. Feature Extraction

We extracted the features by tokenizing the concatenated messages of each user into sequences of up to 512 tokens using the tokenizer in the case **bert-base-spanish-wwm-cased** [22]. The tokenized sequences were then processed through the transformer model to generate 768-dimensional embeddings, obtained by applying mean pooling to the last hidden-state outputs. These embeddings constituted the feature matrices for Tasks 1 and 2, which we directly fed into our classification pipelines. To handle sequences exceeding the 512-token limit, we applied truncation to retain the first 512 tokens while using padding to ensure uniform sequence length.

4.3. Regularization

To mitigate class imbalance in Task 1 and Task 2, we applied random oversampling (using RandomOverSampler with a fixed seed) directly on the 768-dimensional embedding matrices, generating additional examples of minority classes until we reached parity with the majority; the balanced datasets were then split into training (80%) and validation (20%) subsets to support downstream cross-validation and model selection.

4.4. Models Training

This subsection overviews the supervised classifiers in Task 1 (binary risk detection) and Task 2 (multiclass addiction-type classification). We used PyCaret, a machine learning library that facilitates model development and evaluation by automating preprocessing, model training, and algorithm comparison. Additionally, MLflow, an open-source platform for managing the machine learning lifecycle, was integrated to systematically track experiment runs, log model parameters, and ensure reproducibility. We evaluated the classifiers [23] detailed in Table 3 on the balanced 768-dimensional embeddings.

Table 3
Evaluated Classification Models

Abbreviation	Model Name
lr	Logistic Regression
ridge	Ridge Classifier
svm	Support Vector Machine
rf	Random Forest Classifier
lightgbm	Light Gradient Boosting Machine
et	Extra Trees Classifier
gbc	Gradient Boosting Classifier
nb	Naive Bayes
knn	k-Nearest Neighbors Classifier
dt	Decision Tree Classifier
lda	Linear Discriminant Analysis
ada	AdaBoost Classifier
qda	Quadratic Discriminant Analysis
dummy	Dummy Classifier

We trained each model and evaluated using 10-fold cross-validation on 80% of the training data, with Accuracy and F1 score as the primary evaluation metrics. MLflow automatically recorded all experimental configurations and validation results, enabling a systematic and reproducible comparison of classifiers, with the best-performing model selected for final evaluation on the test set.

4.5. Evaluation

At this stage, we conducted a rigorous comparison of the results obtained by the different models for both Task 1 (binary risk detection of gambling disorder) and Task 2 (multiclass classification of gambling type). We selected the best-performing model based on its performance across the metrics: Accuracy, Precision, Recall, F1 score, AUC, Kappa, and Matthews Correlation Coefficient (MCC). These metrics reflect different dimensions of predictive quality and are particularly relevant for evaluating model behavior in class imbalance and uncertainty. By assessing the classifiers against this comprehensive set of criteria, we ensured a rigorous and interpretable comparison of model performance for each task.

4.5.1. Results of Training Process for Binary Classification Task 1

Based on the evaluation results presented in Table 4, LightGBM was selected as the optimal classifier for Task 1 due to its strong performance across key metrics, particularly Accuracy (0.7505), F1 score (0.7315), and MCC (0.5072). The specific hyperparameter configuration for this model is detailed in Table 5.

Table 4
Performance Comparison of Classification Models (Task 1 - Binary Metrics)

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (s)
lightgbm	0.7505	0.8301	0.6909	0.7839	0.7315	0.5005	0.5072	0.1270
lr	0.7426	0.8272	0.7120	0.7601	0.7303	0.4846	0.4906	0.2380
ridge	0.7424	0.8211	0.7026	0.7657	0.7288	0.4852	0.4906	0.0270
et	0.7397	0.8303	0.6798	0.7742	0.7221	0.4785	0.4837	0.0410
rf	0.7289	0.8333	0.6532	0.7721	0.7053	0.4576	0.4650	0.0620
ada	0.7208	0.7743	0.6754	0.7637	0.7088	0.4419	0.4541	0.1500
gbc	0.7179	0.8099	0.6863	0.7392	0.7092	0.4358	0.4396	0.0390
svm	0.7047	0.8249	0.6319	0.7974	0.6549	0.4090	0.4561	0.0250
lda	0.6965	0.6962	0.6588	0.7254	0.6802	0.3928	0.4029	0.0240
knn	0.6935	0.7523	0.6588	0.7088	0.6785	0.3869	0.3907	0.2260
dt	0.6803	0.6801	0.6708	0.7008	0.6778	0.3600	0.3681	0.0280
qda	0.6613	0.7704	0.9728	0.6045	0.7434	0.3214	0.4166	0.0200
nb	0.6532	0.6673	0.4360	0.7586	0.5475	0.3052	0.3352	0.0210
dummy	0.4878	0.5000	0.5000	0.2446	0.3285	0.0000	0.0000	0.0150

Table 5
LightGBM Hyperparameter Configuration for Task 1

Parameter	Value
boosting_type	gbdt
learning_rate	0.1
max_depth	-1
n_estimators	100
num_leaves	31
n_jobs	-1
min_child_samples	20
min_child_weight	0.001
random_state	123

4.5.2. Results of Training Process for Classification Task 2

The results for Task 2 presented in Table 6 demonstrate that Logistic Regression is the top-performing classifier, achieving outstanding metrics with Accuracy (0.9907) and F1 score (0.9906). Table 7 details the hyperparameter configuration for this model to ensure reproducibility.

5. Results of VerbaNexAI in MentalRiskES Task Evaluation

This section outlines the results obtained by the VerbaNexAI team in evaluating the MentalRiskES shared task. The challenge focused on the detection of mental health disorders, with particular attention to the early identification of gambling-related behavior in Spanish-language comments collected from Telegram and Twitch. We used the same model across the three evaluation runs for each subtask: LightGBM for Task 1 and Logistic Regressor for Task 2.

Table 6
Performance Comparison of Classification Models (Task 2 - Macro Metrics)

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
lr	0.9907	0.0000	0.9907	0.9915	0.9906	0.9876	0.9879
ridge	0.9884	0.0000	0.9884	0.9897	0.9883	0.9845	0.9850
svm	0.9837	0.0000	0.9837	0.9847	0.9837	0.9783	0.9786
rf	0.9837	0.9989	0.9989	0.9837	0.9836	0.9783	0.9788
lightgbm	0.9814	0.9996	0.9814	0.9830	0.9813	0.9753	0.9759
et	0.9768	0.9992	0.9992	0.9785	0.9766	0.9690	0.9697
gbc	0.9676	0.0000	0.9676	0.9698	0.9674	0.9568	0.9577
nb	0.9652	0.9926	0.9652	0.9689	0.9653	0.9536	0.9548
knn	0.9445	0.9905	0.9445	0.9519	0.9435	0.9259	0.9286
dt	0.9212	0.9476	0.9212	0.9278	0.9214	0.8950	0.8972
lda	0.8610	0.0000	0.8610	0.8725	0.8606	0.8145	0.8183
ada	0.6551	0.0000	0.6551	0.5971	0.5972	0.5402	0.5741
qda	0.2500	0.0000	0.2500	0.0626	0.1001	0.0000	0.0000
dummy	0.2360	0.5000	0.2360	0.0558	0.0902	0.0000	0.0000

Table 7
Logistic Regression Hyperparameter Configuration for Task 2

Parameter	Value
C	1.0
penalty	l2
solver	lbfgs
max_iter	1000
tol	0.0001
random_state	123

5.1. Task 1: Risk Detection of Gambling Disorders

The VerbaNexAI Lab team approached Task 1 by deploying the LightGBM classifier to detect early signs of gambling behavior in Telegram messages. Table 8 presents the classification results for Task 1, including Accuracy, Macro, and Micro averaged Precision, Recall, and F1 scores. In this subtask, the model achieved an overall Accuracy of 0.519, with a macro-F1 of 0.342, highlighting moderate performance across classes. The Micro-F1 score reached 0.519, indicating consistent per-instance prediction performance.

In addition to classification metrics, we evaluated the system's efficiency in timely detection using ERDE5, ERDE30, latencyTP, detection speed, and latency-weighted F1. As shown in Table 9, VerbaNexAI Lab achieved the lowest ERDE5 score (0.274) among all participating teams, demonstrating its strong capability for early detection of gambling-related risks. The team also obtained a competitive ERDE30 (0.250), latencyTP of 2, and a top detection speed of 0.990, resulting in a latency-weighted F1 score of 0.677. While classification performance suggests room for improvement in class balance, the system's outstanding timeliness metrics confirm that it is highly effective for real-time monitoring scenarios.

Table 8
Classification-based evaluation in Task 1 – VerbaNexAI Lab

Rank	Run	Accuracy	Macro_P	Macro_R	Macro_F1	Micro_P	Micro_R	Micro_F1
27, 28, 29	0, 1, 2	0.519	0.259	0.500	0.342	0.519	0.519	0.519

Table 9
Latency-based evaluation in Task 1 – VerbaNexAI Lab

Rank	Run	ERDE5	ERDE30	LatencyTP	Speed	Latency-weightedF1
27, 28, 29	0, 1, 2	0.274	0.250	2	0.990	0.677

5.2. Task 2: Type of Addiction Detection

Task 2 focused on a multiclass classification challenge to identify each user's specific type of gambling based on their message history. The VerbaNexAI Lab team used a Logistic Regression model for all three submitted runs. The evaluation considered the assigned label's correctness and the prediction's timeliness.

The model achieved an Accuracy of 0.813, along with a Macro Precision of 0.846, Macro Recall of 0.769, and Macro F1 score of 0.780, as shown in Table 10. These metrics indicate a well-balanced performance across the different classes. Furthermore, the Micro metrics (Precision, Recall, and F1) were also consistent, each scoring 0.813, which suggests uniform instance-level classification.

These results highlight the effectiveness of the Logistic Regression model in accurately distinguishing between different addiction types, delivering solid performance in a multiclass classification setting.

Table 10
Classification-based evaluation in Task 2 – VerbaNexAI Lab

Rank	Run	Accuracy	Macro_P	Macro_R	Macro_F1	Micro_P	Micro_R	Micro_F1
13, 14, 15	0, 1, 2	0.813	0.846	0.769	0.780	0.813	0.813	0.813

6. Error Analysis

To evaluate the limitations of our models, we conducted an analysis of misclassified cases using the official gold labels published for the test set. This analysis reveals distinctive error patterns that provide insights into the limitations and strengths of our approach.

The analysis identified three primary sources of error affecting both models: the inability to capture subtle mood and emotional changes that are characteristic of pathological gambling behaviors, insufficient message context for users with limited posting history, and the presence of gambling-specific terminology that appears across multiple categories without adequate contextual understanding. Additionally, confusions were observed between categories with overlapping vocabulary, particularly between Trading and Betting due to shared financial terminology.

However, the data augmentation strategy for the Lootboxes category proved effective, significantly improving its classification despite the original class imbalance. The error patterns suggest the need to focus on incorporating temporal emotional modeling and developing domain-specific lexicons for Spanish gambling terminology.

7. Carbon Emission

We tracked resource consumption and emissions during the complete pipeline execution using CodeCarbon API [24] to assess the computational requirements and environmental impact of our approach. Understanding these metrics is crucial for identifying approaches suitable for deployment on personal computers or resource-constrained environments. Table 11 details our computational hardware configuration used for the experiments, while Table 12 presents the energy consumption and carbon emissions recorded during the complete pipeline execution.

Table 11
Computational Hardware Configuration

Component	Specification
CPU	12th Gen Intel(R) Core(TM) i7-12700KF (20 cores)
GPU	1 x NVIDIA GeForce RTX 4060 Ti
RAM	31.84 GB
Country ISO Code	COL

Table 12
Energy Consumption and Carbon Emissions

Metric	Value
GPU Energy	0.095 kWh
RAM Energy	0.002 kWh
CPU Energy	0.195 kWh
Total Energy Consumed	**0.247 kWh**
CO_2 Emissions	**0.065 kg**

The results demonstrate that our approach maintains reasonable computational efficiency for resource-constrained environments, with the majority of energy consumption occurring during the embedding generation phase using the Spanish BERT model, while downstream classification tasks required minimal additional resources.

8. Conclusions and Future Work

This paper presented VerbaNexAI's approach to the MentalRiskES 2025 shared task for early detection of gambling disorders in Spanish social media content. Our methodology combined rigorous text preprocessing, contextual embeddings from Spanish BERT models, class balancing through random oversampling, and systematic model selection.

For Task 1 (risk detection), LightGBM achieved an accuracy of 0.7505 and an F1 score of 0.7315 during cross-validation. At the same time, Logistic Regression demonstrated exceptional performance in Task 2 (addiction type classification) with 0.9907 accuracy. However, when evaluated on the official test set, our system reached 0.519 accuracy and a Macro-F1 of 0.342 in Task 1, alongside more robust results in Task 2 with 0.813 accuracy and a Macro-F1 of 0.780.

Notably, our approach demonstrated robust early detection performance, securing the lowest ERDE5 score (0.274) among all participating teams. It also had an impressive detection speed of 0.990 and a latency-weighted F1 of 0.677. These metrics highlight the system's effectiveness for real-time monitoring scenarios.

This work's primary contribution demonstrates that transformer-based embeddings combined with traditional machine learning algorithms can detect early signs of gambling disorders in Spanish text. Our approach prioritized detection timeliness, crucial for real-world mental health monitoring, where early intervention significantly improves outcomes. The outstanding early detection metrics validate the system's practical potential for mental health risk monitoring. At the same time, our class imbalance handling techniques proved effective, particularly for Task 2's underrepresented Lootboxes class.

The noticeable difference between cross-validation performance and official evaluation results in Task 1 highlights the challenge of generalizing mental health risk detection across diverse user populations and communication contexts. Our preprocessing approach may have overlooked certain language nuances specific to Spanish-speaking gambling communities, particularly evolving slang or context-dependent expressions used in Telegram and Twitch platforms.

Future research should incorporate temporal modeling techniques to better capture gambling behavior progression, develop domain-specific lexicons for Spanish gambling terminology, combine text analysis with behavioral metrics (message frequency, time patterns), and investigate explainable AI techniques to enhance transparency for mental health professionals. These improvements would strengthen early detection systems for mental health risks in Spanish-language communities, supporting timely interventions for vulnerable individuals.

Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnológica de Bolívar (UTB) in Cartagena, Colombia.

Declaration on Generative AI

During the preparation of this investigation, Claude Sonnet 4 was used for the revision of translations into English, as well as for grammatical and spelling correction. After using this tool, the content was reviewed and edited as necessary, and full responsibility for the content of the publication is assumed.

References

[1] World Health Organization, Who special initiative for mental health, https://www.who.int/initiatives/who-special-initiative-for-mental-health, 2025. Accessed: 2025-05-04.

[2] World Health Organization, Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide, https://n9.cl/wesf4, 2022. Accessed: 2025-05-04.

[3] World Health Organization, World Report on Universal Health Coverage, Technical Report 9789240026643, World Health Organization, 2022. Accessed: 2025-05-04.

[4] H. Wardle, L. Degenhardt, V. Marionneau, G. Reith, C. Livingstone, M. Sparrow, L. T. Tran, B. Biggar, C. Bunn, M. Farrell, V. Kesaite, V. Poznyak, J. Quan, J. Rehm, A. Rintoul, M. Sharma, J. Shiffman, K. Siste, D. Ukhova, R. Volberg, J. S. Yendork, S. Saxena, The *lancet public health* commission on gambling, The Lancet Public Health 9 (2024) e950–e994. URL: https://doi.org/10.1016/S2468-2667(24)00167-1. doi:10.1016/S2468-2667(24)00167-1.

[5] National Center for Biotechnology Information, Table 3.39: Prevalence of mental disorders by who region, in: Mental Health Competency Frameworks: A Global Perspective, National Academies Press (US), 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK519704/table/ch3.t39, accessed: 2025-05-04.

[6] J. H. Kristensen, S. Pallesen, J. Bauer, T. Leino, M. D. Griffiths, E. K. Erevik, Suicidality among individuals with gambling problems: A meta-analytic literature review, Psychol Bull 150 (2024) 82–106. doi:10.1037/bul0000411.

[7] S. Hautamäki, V. Marionneau, S. Castrén, J. Palomäki, S. Raisamo, T. Lintonen, P. Pörtfors, T. Latvala, Methodologies and estimates of social costs of gambling: A scoping review, Social Science Medicine 371 (2025) 117940. doi:https://doi.org/10.1016/j.socscimed.2025.117940.

[8] R. Bijker, N. Booth, S. S. Merkouris, N. A. Dowling, S. N. Rodda, Global prevalence of help-seeking for problem gambling: A systematic review and meta-analysis, Addiction 117 (2022) 2972–2985. doi:10.1111/add.15952.

[9] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: CLEF 2023: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer-Verlag, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[10] 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.

[11] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Raéz, Overview of MentalRiskES at IberLEF 2023:

Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 71 (2023) 329–350. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[12] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalRiskES at IberLEF 2024: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 73 (2024) 435–448. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6629.

[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] A. Molina, X. Huang, L.-F. Hurtado, F. Pla, Elirf-upv at eRisk 2023: Early detection of pathological gambling using svm, CEUR-WS 3497 (2023) 736–742. URL: https://ceur-ws.org/Vol-3497/paper-062.pdf.

[15] H. Fabregat, A. Duque, L. Araujo, J. Martinez-Romo, Nlp-uned-2 at eRisk 2023: Detecting pathological gambling in social media through dataset relabeling and neural networks, CEUR-WS 3497 (2023) 672–683. URL: https://ceur-ws.org/Vol-3497/paper-056.pdf/.

[16] A. M. Mármol-Romero, F. M. Plaza-Del-Arco, A. Montejo-Ráez, Sinai at eRisk@CLEF 2023: Approaching early detection of gambling with natural language processing, CEUR-WS 3497 (2023) 743–751. URL: https://ceur-ws.org/Vol-3497/paper-063.pdf.

[17] H. Thompson, L. Cagnina, M. Errecalde, Strategies to harness the transformers' potential: Unsl at eRisk 2023, CEUR-WS 3497 (2023) 791–804. URL: https://ceur-ws.org/Vol-3497/paper-068.pdf.

[18] H. Fabregat, A. Duque, L. Araujo, J. Martinez-Romo, Uned-nlp at eRisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors, CEUR-WS 3180 (2022) 894–904. URL: https://ceur-ws.org/Vol-3180/paper-71.pdf.

[19] B. Perrot, J. B. Hardouin, E. Thiabaud, A. Saillard, M. Grall-Bronnec, G. Challet-Bouju, Development and validation of a prediction model for online gambling problems based on players' account data, Journal of behavioral addictions 11 (2022) 874–889. URL: https://pubmed.ncbi.nlm.nih.gov/36125924/. doi:10.1556/2006.2022.00063.

[20] Álvarez Ojeda Pablo, C.-R. M. Victoria, S. Anastasia, M.-R. Arturo, The precom-sm corpus: Gambling in spanish social media, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 17–28.

[21] J. Cuadrado, E. Martinez, J. Cuadrado, J. C. Martinez-Santos, E. Puertas, Verbanex ai at dipromats 2024: Enhancing propaganda detection in diplomatic tweets with fine tuned bert and integrated nlp techniques, CEUR-WS 3756 (2024). URL: https://ceur-ws.org/Vol-3756/{MentalRiskES}2024_paper8.pdf.

[22] Cañete, José, Chaperon, Gabriel, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[23] E. Martinez, J. Cuadrado, J. C. Martinez-Santos, E. Puertas, Automated detection of depression and anxiety using lexical and phonestheme features in spanish texts, CEUR-WS 3756 (2024). URL: https://ceur-ws.org/Vol-3756/{MentalRiskES}2024_paper8.pdf.

[24] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, L. O. N. de Araújo, JPW, MinervaBooks, mlco2/codecarbon: v2.4.1, 2024. URL: https://doi.org/10.5281/zenodo.11171501. doi:10.5281/zenodo.11171501.