Early Detection of Gambling Addiction Risk in Spanish: purple_john at MentalRiskES2025

Cristina Damov[1,*,†], Adina Ion[1,†] and Crystal Schlupek[1,†]

[1]University of Bucharest, St. Academiei 14, Bucharest, Romania

Abstract

This paper presents our submission to the MentalRiskES 2025 shared task, focused on the early detection of gambling addiction risk from Spanish-language social media posts. Our approach explored a range of deep learning and traditional machine learning models, including CNN, MobileNet, Inception, and transformer-based architectures such as RoBERTa and BETO, combined with pre-trained embeddings like GloVe and FastText. Despite extensive experimentation and model diversity, classification performance remained modest, with our best accuracy reaching only 51% on final tests. These results suggest potential limitations not only in model capacity but also in the dataset and task formulation, where user-level annotations and subtle language differences may be insufficient to reliably distinguish risk levels. Further improvements are needed to better detect behavioral risk in informal and multilingual data.

Keywords

mental health, NLP, Spanish, social media

1. Introduction

Mental health concerns have become increasingly visible in recent years, with global estimates from the World Health Organization indicating that one in eight people is affected by a mental disorder [4]. The COVID-19 pandemic further exacerbated many of these issues, including anxiety, depression, and various forms of addiction [5]. As people increasingly turn to social media to share their thoughts and experiences, these platforms have become a valuable source of information for identifying early signs of mental health risks.

While much existing work in the field of computational mental health has focused on conditions like depression or suicidal ideation, gambling addiction remains relatively underexamined, particularly in non-English languages. Gambling disorder is often difficult to detect, as its linguistic markers are subtler and more context-dependent compared to other mental health conditions.

This paper presents our submission to the third edition of a shared task, focused on early detection of mental health risks in Spanish social media, organized as part of IberLEF 2025 [1]. This year's task introduced two new subtasks: one centered specifically on detecting gambling disorder risk, and another on identifying types of addiction. Unlike previous editions, which focused more broadly on psychological risk, this year's theme required participants to address the nuances of addictive behaviors in a Spanish-language context.

Our goal was to develop an NLP system capable of identifying high risk of developing a gambling addiction from a stream of user posts. In doing so, we aimed not only to contribute to the technical challenge but also to support the broader goal of enabling earlier and more targeted mental health interventions, especially in underrepresented languages and conditions.

IberLEF 2025 September 2025, Zaragoza, Spain

*Corresponding author.

damovcristina@gmail.com (C. Damov); adinaion53@gmail.com (A. Ion); crystalschlupek4@gmail.com (C. Schlupek)

0009-0002-2580-7284 (C. Damov)

2. Task

2.1. Data

Our work addressed Subtask 1: Risk Detection of Gambling Disorders [6] , which involved classifying messages as suggesting either a high risk or a low risk of the user developing a gambling addiction. The challenge was framed as a binary classification problem, where systems had to make this determination based on a user's timeline of messages. Evaluation metrics included not just prediction scores, but also computational resources (duration, emissions, energy consumed etc.)
The data provided consisted of Spanish-language social media messages, sent on Telegram groups or Twitch chats. These annotations posed additional challenges due to the informal nature of the text, potential noise, and the subtle linguistic cues that may signal gambling-related behavior. This subtask aimed to push forward the use of NLP tools for behavioral risk assessment, particularly in less-resourced languages and nuanced domains like gambling addiction.

2.2. Submission Procedure

The predictions were evaluated in an online, round-based manner. The program setup included server connection and API calls to receive a batch of messages from users and submit a prediction for each one before proceeding to the next round. This workflow simulated real-time detection and increased the challenge, as any message could be the user's last.
Notably, true labels were set at user level, however predictions were made at message level. This meant models were also challenged by the fact that some messages were not relevant to the user's risk label and there was no user message history to predict contextually.

3. Methodology

3.1. Preprocessing

The quality of the input data has a significant impact on the effectiveness and accuracy of text classification models and sentiment analysis in the field of natural language processing (NLP). Several data pretreatment procedures were tested to enhance the accuracy and consistency of sentiment analysis prediction in binary text categorization.

Lowercasing
Converting all text to lowercase offers several benefits, including text normalization, vocabulary reduction, compatibility with word embedding models, and ensuring consistency in text processing pipelines. This standardizes the text and ensures that the model does not treat words with different cases as distinct entities. While this step improves consistency, it may also lead to a loss of information from capitalized proper nouns or acronyms.

Punctuation Removal
Punctuation marks, though important for grammatical structure, generally do not add semantic value in this task. Removing punctuation reduces noise and simplifies tokenization, allowing the model to focus on word-level content without distraction from syntactic elements.

Lemmatization
Lemmatization reduces words to their dictionary base form, decreasing word variations to a single representative. This reduces feature sparsity and emphasizes core meanings. Unlike stemming, lemmatization considers part of speech and context, making it more linguistically accurate.

Stemming
Stemming is another normalization technique that removes suffixes to reduce words to their root form. It is a faster, simpler alternative to lemmatization, but may produce non-standard root forms and overlook

contextual meaning. Despite its crudeness, stemming can be effective in reducing dimensionality when linguistic precision is not critical.

Stop Words Removal
Stop words are common words that occur frequently in a language. These words do not carry significant sentiment value and can be safely removed from the text, reducing noise in the dataset and making it more focused and meaningful for sentiment analysis.

3.2. Additional data used

No external or augmented data was used. The model was trained strictly on the dataset provided in the shared task. This was a conscious choice, as most available mental health datasets focus on issues like depression or anxiety and don't align well with the specific language and behaviors related to gambling addiction. We prioritized maintaining domain relevance and consistency with the task's objectives, even if it limited the amount of training data available.

3.3. Advanced Natural Language Processing Models

To improve early detection of mental health conditions in Spanish social media comments, several advanced NLP models were implemented. These models capture different linguistic patterns and contextual nuances, each contributing to the classification process. The models used are:

- BiLSTM: Applied for capturing bidirectional context in text, which is valuable for detecting subtle emotional expressions.
- CNN with FastText [7]: Used to identify local linguistic patterns. This combination was effective in processing short, informal messages.
- MobileNet [8](Adapted for Text): A lightweight deep neural network adapted for text classification using word/document embeddings. It achieved strong accuracy (0.73) while maintaining low computational cost, making it suitable for real-time applications.
- Transformer-based Models (RoBERTa [9] and BETO [10]): These models were evaluated for their deep contextual embeddings. Despite their potential, they achieved less than 60% accuracy in this setting, likely due to domain mismatch and limited fine-tuning.
- LightGBM with FastText: This gradient boosting model effectively handled high-dimensional text features and performed better than simpler models like logistic regression in several configurations.
- Logistic Regression: Used as a baseline for comparison. When paired with quality embeddings (e.g. FastText), it showed competitive results, validating its role as a strong, interpretable baseline.

3.4. Classifiers and Ensemble Techniques

To enhance performance, various classifiers and ensemble methods were applied. These ranged from lightweight models to more complex deep learning architectures, offering a broad perspective on their trade-offs:

- Logistic Regression with Embeddings: Served as a strong baseline when used with semantic-rich embeddings like FastText and GloVe [11]. It was also useful in ensemble strategies due to its probabilistic output.
- LightGBM: Demonstrated robustness and efficiency, especially in handling class imbalance. It integrated well with FastText features and benefited from histogram-based optimization for faster training.
- FastText + CNN: This configuration effectively captured local patterns indicative of mental distress and provided interpretable, efficient classification.
- FastText + Logistic Regression: A computationally efficient strategy that maintained stable and reliable performance, especially in resource-constrained settings.

- GloVe-Based Models: Both English and Spanish embeddings were tested, with English GloVe surprisingly outperforming Spanish in certain cases, possibly due to language mixing in social media text.
- MobileNet with Dimensionality Reduction: This setup delivered one of the best trade-offs between accuracy (0.73) and speed. It showed promise for applying transfer learning techniques from vision models to NLP tasks.
- Transformer Models (RoBERTa and BETO): Although powerful, these models underperformed (<60% accuracy on validation data), likely due to insufficient fine-tuning and mismatch with the target domain. Their computational cost also limited practical deployment in this project.

4. Results and Discussion

Each model was evaluated using accuracy, F1 score and early detection performance, as summarized in Table 1:

- BERT was initially expected to deliver the best results given its former state-of-the-art status and good results in other studies [12][13], but in practice its performance suffered under tight resource constraints. It required substantial training time and memory, leading us to remove it from further consideration.
- GloVe embeddings were tested in both English and Spanish. The English models outperformed the Spanish ones, likely due to superior slang and informal-speech coverage in the English vectors.
- BiLSTM, as a powerful recurrent architecture, achieved solid effectiveness but at the cost of high computational and memory demands when setting trainable=True.
- MobileNet proved to be fast, lightweight, and surprisingly accurate, making it well suited for resource-limited environments, as illustrated by its confusion matrix in Figure 1.
- Inception achieved performance and inference speed on par with MobileNet, making it a reliable choice among the tested models (Figure 2).

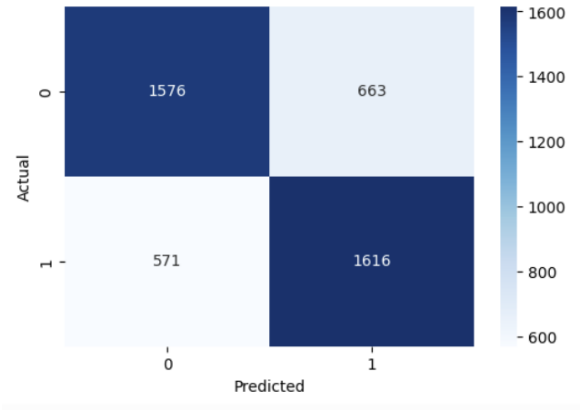

Figure 1: MobileNet confusion matrix

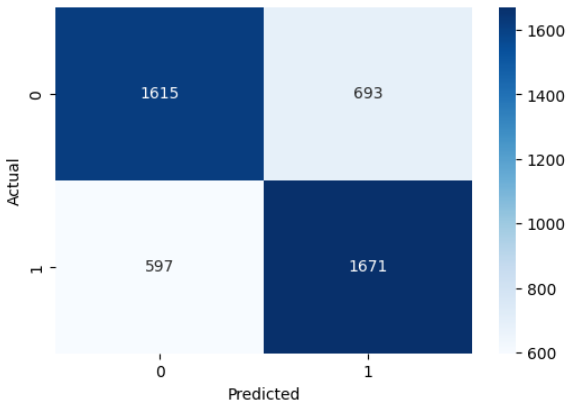

Figure 2: Inception confusion matrix

Applying dimensionality reduction dramatically reduced training time (especially for BERT) but did not yield any appreciable gains in accuracy.

4.1. Hyperparameters and Training Configuration

We experimented with a set of standard hyperparameters. Minor tuning was performed using manual adjustments and a grid search over key parameters, such as learning rate and batch size, to optimize performance across different models. The goal was to balance training stability, speed, and generalization while staying within computational constraints.
The final configuration was:

Table 1
Comparative performance of models

Model	Precision (macro avg)	Recall (macro avg)	F1-score (macro avg)	Accuracy
RoBERTa	0.59	0.59	0.58	0.5736
BETO	0.57	0.57	0.56	0.5680
MobileNet	0.72	0.72	0.72	0.7181
Inception	0.73	0.73	0.73	0.7291

- Epochs: 10-25
  Training was capped at 25 epochs, with early stopping based on F1 score to prevent overfitting. Most models converged within 15-20 epochs.
- Batch size: 16-64
  Smaller batch sizes (16) worked well for models with high memory use (e.g., BiLSTM), while 64 was used for lighter models like Logistic Regression or CNN.
- Learning rate: 0.001
  A learning rate of 0.001 was stable across most models (especially CNN and MobileNet). Higher rates led to unstable training, while lower ones often did not converge.
- Optimizer: Adam
  Adam was used for all deep learning models due to its adaptability and efficient convergence on sparse data.
- Embedding dimension: 300
  Both GloVe and FastText embeddings were used with a 300-dimensional vector size. This offered a good balance between expressiveness and computational cost.
- Maximum sequence length: 100 tokens
  Limiting input length helped control memory usage and avoided overfitting to noisy inputs. Most posts were short, so truncation rarely lost meaningful context.
- Early stopping: Enabled
  Training halted when the validation F1 score failed to improve for several epochs, helping avoid overfitting, especially on smaller or noisy subsets.
- Validation split: 20%
  A fifth of the training data was reserved for validation. No external datasets were used to ensure consistency with the task guidelines.

4.2. Final Model Selection

Some preprocessing steps, like lemmatization, were tested early on but later excluded from the final pipeline. They slowed down text processing and didn't lead to any improvement in accuracy. Since they added extra complexity without offering real benefits, we chose to leave them out and stick to a simpler setup that was faster and more efficient (lowercasing, punctuation and stopword removal).
Although multiple embeddings were tested, including FastText and GloVe, both English and Spanish-specific versions, we ultimately kept only the English GloVe embeddings in the final models. Surprisingly, they consistently outperformed the Spanish ones, likely due to better coverage of informal language and cross-lingual slang often used in the dataset. This choice also helped simplify the pipeline and reduced training time without hurting accuracy.
The comparative performance of the final models on validation data is presented in Table 2.

5. Error Analysis

Despite using various filters and preprocessing steps (stopword removal, punctuation stripping, lowercasing, lemmatization), data issues persisted. The dataset was annotated at user level, which caused

Table 2
Comparative performance of final models

Model	Precision (macro avg)	Recall (macro avg)	F1-Score (macro avg)	Accuracy
MobileNet	0.72	0.72	0.72	0.7181
Inception	0.72	0.72	0.72	0.7212
Sequential	0.67	0.67	0.66	0.6661

mismatches, for example, meaningless messages like “Si” (“Yes”) were marked as risky. This limited model learning and introduced noise.
A detailed review of failed predictions revealed several causes:

- Noisy inputs: messages like "jajaja" (“hahaha”), emojis, or very short comments were incorrectly labeled as risky, affecting training quality and creating a model prone to many false positive labels. This is also reflected in the fact that several models have lower precision compared to their F1 scores.
- Weak lexical distinction: no strong trends or word-level differences between risk 0 and risk 1, making separation difficult.
- Linguistic limitations: use of slang, regional Spanish, or code-switching (ES/EN) reduced embedding effectiveness.
- Domain-specific overlap: both risk and non-risk users share the same community slang and terminology (betting or crypto terms), reducing lexical contrast typically used for classification.

The comparative analysis of messages labeled risk 0 versus risk 1 revealed no meaningful differences. Although Figure 3 (hourly frequency plot) and Figure 4 (emotion-count chart) display certain trends, such as peaks of activity at specific hours and a predominance of anticipation and negative emotions (which is unsurprising in a gambling context); these patterns do not clearly separate the two risk groups. Likewise, simple metrics like word count and other extracted features fail to distinguish risk 0 from risk 1 in any reliable way.

6. Resource Impact and Emissions

In addition to evaluating classification performance, it is increasingly important to consider the computational and environmental cost of training and deploying NLP models. This is especially relevant in shared tasks and evaluation campaigns where efficiency and sustainability are valued alongside predictive accuracy.
Our submissions included three distinct models, each trained and evaluated independently, yet all sharing a common emphasis on energy efficiency and hardware-conscious design. The results for these three runs and baseline models, including F1 score, emissions, and CPU usage, are summarized in Table 3.

Table 3
Score and emissions for final models on testing data

Model	Run Index	Precision (macro)	F1 Score (macro)	Duration Mean	Emissions Mean	CPU Energ
Robertuito		0.657	0.428			
MobileNet	0	0.509	0.519	2.081	3.40×10^{-5}	$6.94 \times$
Inception	1	0.259	0.519	2.082	3.40×10^{-5}	$6.94 \times$
Roberta Base		0.259	0.342			
Sequential	2	0.258	0.513	2.083	3.40×10^{-5}	$6.94 \times$

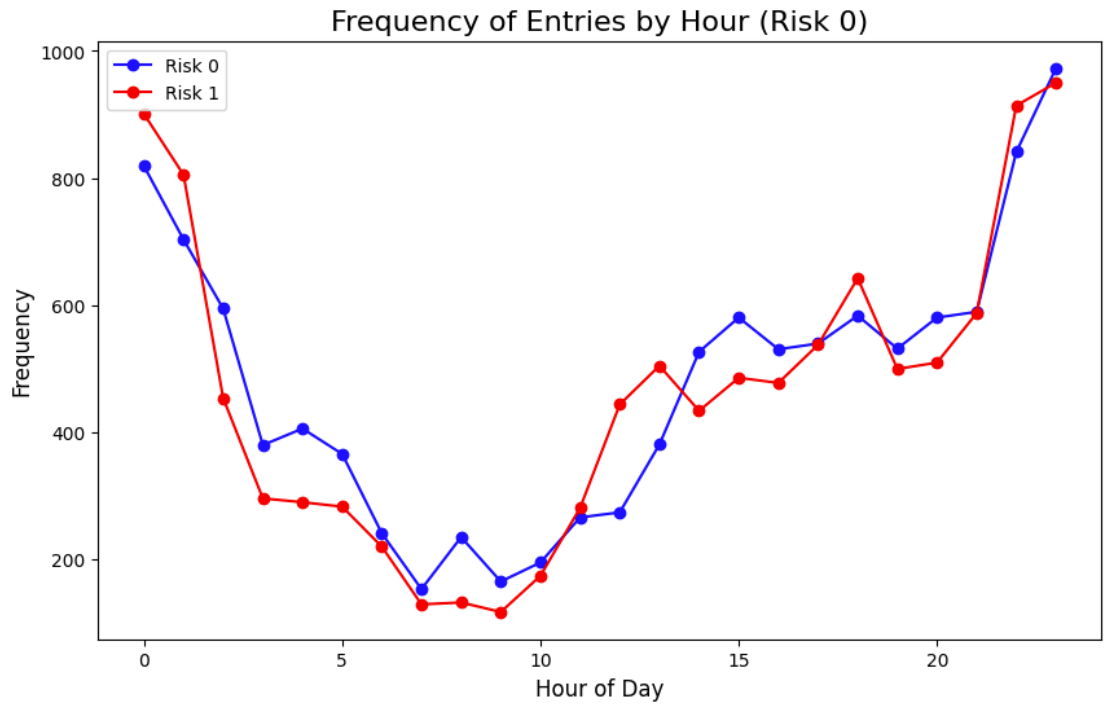

Figure 3: Hourly frequency plot split between risk classes.

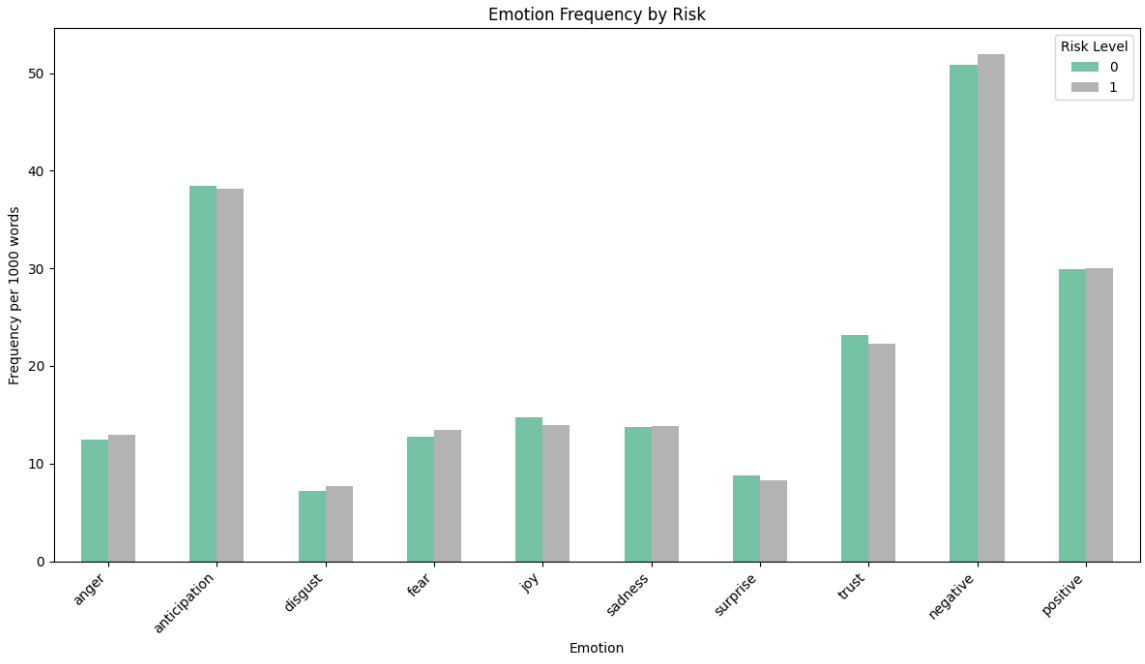

Figure 4: Emotion-count chart split between risk classes.

These submissions reflect the team's consistent focus on resource-conscious model design. Although transformer-based architectures like RoBERTa and BETO were evaluated, they were ultimately excluded due to modest performance and high resource demands.

While the accuracy of these models (51-52%) did not reach the top rankings, the energy consumed was of lower magnitude than that of GPU-based or transformer-based solutions. Duration and emissions were similar between the runs. This highlights a crucial trade-off in applied machine learning: models

that are sustainable and efficient may be more viable for real-world deployment, especially in resource-constrained environments.

7. Conclusions

This study, carried out as part of the MentalRiskES 2025 shared task, underlined the difficulty of detecting gambling addiction risk from short, informal social media messages in Spanish. Despite testing a wide range of models, including traditional classifiers, CNNs, MobileNet, and transformers, performance remained modest, with limited distinction between high-risk and low-risk users. MobileNet and Inception achieved the best trade-off between speed and accuracy, while transformer models underperformed in this domain. Much of the challenge stemmed from noisy data, inconsistent labeling, and a lack of strong linguistic cues to separate classes.
While the results were not especially strong, they reflect the complexity of the task and the need for better data, clearer risk definitions, and more context-aware approaches. Future work should focus on these aspects to improve real-world applicability.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly and ChatGPT in order to: paraphrase, identify and correct typos and grammatical mistakes. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Mármol-Romero, A. M., Álvarez-Ojeda, P., Moreno-Muñoz, A., Plaza-del-Arco, F. M., Molina-González, M. D., Martín-Valdivia, M. T., Ureña-López, L. A., & Montejo-Ráez, A. (2025). Overview of MentalRiskES at IberLEF 2025: Early Detection of Mental Disorders Risk in Spanish. *Procesamiento del Lenguaje Natural*, 75.

[2] Álvarez-Ojeda, P., Cantero-Romero, M. V., Semikozova, A., & Montejo-Ráez, A. (2025). The PRECOM-SM corpus: Gambling in Spanish social media. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 17–28).

[3] González-Barba, J. A., Chiruzzo, L., & Jiménez-Zafra, S. M. (2025). 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.

[4] World Health Organization. (2022, June 8). Mental disorders. Retrieved from https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[5] World Health Organization. (2020, October 5). COVID-19 disrupting mental health services in most countries, WHO survey. Retrieved from https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

[6] Montejo-Ráez, A., Mármol-Romero, A. M., Álvarez-Ojeda, P., Plaza-del-Arco, F. M., Ureña-López, L. A., Martín-Valdivia, M. T., Molina-González, M. D., & Moreno-Muñoz, A. (2025). MentalRiskES 2025: Early detection of mental disorders risk in Spanish. *IberLEF 2025*. Retrieved May 17, 2025, from https://sites.google.com/view/mentalriskes2025/tasks

[7] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications [Technical report]. *arXiv*. https://arxiv.org/abs/1704.04861

[9] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach [Technical report]. *arXiv*. https://arxiv.org/abs/1907.11692

[10] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Practical Machine Learning for Developing Countries Workshop at ICLR 2020 (PML4DC)*.

[11] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://aclanthology.org/D14-1162

[12] Smith, E., Peters, J., & Reiter, N. (2024). Automatic detection of problem-gambling signs from online texts using large language models. *PLOS Digital Health*, 3(9), e0000605. https://doi.org/10.1371/journal.pdig.0000605

[13] Bucur, A.-M., Cosma, A., & Dinu, L. P. (2021). Early risk detection of pathological gambling, self-harm and depression using BERT. *arXiv*. https://doi.org/10.48550/arXiv.2106.16175

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