ELiRF-UPV at MentalRiskES 2025: Spanish Longformer for Early Detection of Gambling Addiction Risk

Andreu Casamayor[1,*], Vicent Ahuir[1], Antonio Molina[1] and Lluís-Felip Hurtado[1]

[1] *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia. Spain*

Abstract

This paper describes the approaches the ELiRF-VRAIN team took in the shared tasks of MentalRiskES at IberLEF 2025. The tasks focus on the detection of mental health disorders in Spanish-language social media, specifically: Risk Detection of Gambling Disorders and Type of Addiction Detection. We have developed three approaches: one based on Support Vector Machines, and two based on Transformer architectures, RoBERTa and Longformer. For the Transformer models, we continued pre-training the base models to adapt them to the mental health domain, resulting in two models specifically tailored for this area. During the fine-tuning phase, we applied a data augmentation process using the data provided by the organizing entity. According to the results obtained, our approaches align well with the objectives of the tasks.

Keywords

Longformer, Transformers, Support Vector Machine, Mental disorder detection

1. Introduction

Mental health conditions, including depression, anxiety, and schizophrenia, have become critical global issues, affecting millions of people worldwide. According to the World Health Organization (WHO), mental disorders are characterized by clinically significant disruptions in an individual's thinking, emotional control, or behavior [1].

Alarmingly, nearly one in eight individuals worldwide has a mental illness, with a significant proportion of cases remaining undiagnosed and untreated [1]. Despite growing awareness, the prevalence of mental health disorders continues to rise, and stigma and discrimination toward affected individuals persist. Governments are investing in prevention and treatment initiatives, but the human and material resources shortage limits access to adequate care for many. Furthermore, early detection of mental disorders remains a significant challenge.

Early detection of mental illnesses is therefore essential to improving individuals' lives and reducing their impact on society. Significant progress has been made in automatic detection through the analysis of social media text. However, numerous challenges still hinder this task, including data quality, quantity, and availability. The goal of the MentalRiskES shared tasks is to provide high-quality labeled data in Spanish and to encourage the development of models for the early detection of mental health disorders.

In the 2025 edition [2, 3], the competition consisted of two tasks: (1) Risk Detection of Gambling Disorders and (2) Type of Addiction Detection. Our team participated in both tasks.

We considered three different approaches.

1. The first approach is based on a classical machine learning algorithm: Support Vector Machines (SVM). SVMs have shown reliable performance in long-text classification tasks, making them

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

ancase3@upv.es (A. Casamayor); vahuir@dsic.upv.es (V. Ahuir); amolina@dsic.upv.es (A. Molina); lhurtado@dsic.upv.es (L. Hurtado)

https://vrain.upv.es/elirf/ (A. Casamayor); https://vrain.upv.es/elirf/ (V. Ahuir); https://vrain.upv.es/elirf/ (A. Molina); https://vrain.upv.es/elirf/ (L. Hurtado)

0009-0003-6000-3828 (A. Casamayor); 0000-0001-5636-651X (V. Ahuir); 0000-0001-6537-8803 (A. Molina); 0000-0002-1877-0455 (L. Hurtado)

suitable for this scenario. This approach serves as a baseline to evaluate the effectiveness of traditional models.

2. The second approach leverages a Transformer-based model [4]. Specifically, we employ a pre-trained RoBERTa model [5] as the foundation and perform fine-tuning to adapt it to the task-specific domain. For the fine-tuning process, we consider two distinct datasets: the official dataset provided by the organizers and an expanded version obtained through data augmentation techniques.
3. The final approach follows a similar strategy to the second one; however, to capture a broader context, we utilize a pre-trained Longformer model [6]. The model can leverage more contextual information thanks to its ability to process longer input sequences. For the fine-tuning phase, we use the same datasets as in the previous approach

For Tasks 1 and 2, we submitted three runs, one for each approach described above. The best-performing model was selected in each run through a preliminary evaluation phase, where we tested multiple model configurations and datasets.

2. Description of Dataset and Tasks

The datasets provided by the organizers consisted of a collection of messages sent to various public Telegram groups and Twitch [7]. These groups are characterized by their focus on gambling, and all communication is in Spanish. User labeling was performed through manual and semi-automatic annotation, where users were classified based on the presence or absence of signs of problematic behavior related to gambling addiction.

This year, the focus is on gambling addiction. All users included in the dataset exhibit signs or symptoms of this disorder. The dataset is distributed as follows: 7 users for the trial set, 350 for training, and 160 for testing. It is the same dataset for both tasks; the only aspect that varies is the label assigned to the user, but the content is the same.

The main objective of this competition is to predict mental disorders as early as possible. To simulate a realistic scenario, the organizers implemented a server-based setup that delivers data in packets, each containing one message per user. The system states the prediction for each user based on the current and previously received messages before the next packet arrives. The ultimate goal is to identify the presence of a mental disorder, if any, as early as possible in the message stream.

2.1. Task 1: Risk Detection of Gambling Disorders

Task 1 is a binary classification task aimed at predicting whether users are at high or low risk of developing gambling addiction.

Table 1 shows the distribution among the different labels in the dataset for the first task.

	Train	**Trial**	**Total**
Low risk	178	4	182
High risk	172	3	175
Total	350	7	357

Table 1
Distribution of samples across the Train and Trial partitions of the Task 1 dataset.

To maximize the number of available samples for training, we combined the Train and Trial partitions. The *Total* column in Table 1 displays the final distribution of samples in our training dataset.

2.2. Task 2: Type of Addiction Detection

Task 2 is similar to Task 1; however, for all users, regardless of whether they are at low or high risk, the objective is to predict the specific type of addiction they exhibit. This task is therefore framed as a multi-class classification problem, with the following categories: Betting, Online Gaming, Trading and Crypto, and Loot Boxes. Every user exhibits one addiction, whether this addiction is in a low-risk or high-risk state.

The label distribution in this total dataset can be seen in Table 2.

	Train	Trial	Total
Betting	85	2	87
Online Gaming	104	2	106
Trading and Crypto	135	2	137
Loot Boxes	26	1	27
Total	350	7	357

Table 2
Distribution of samples across the Task 2 dataset.

3. System architecture and Techniques

For this competition, we aimed to investigate three essential factors for this type of task: the size of the context, domain-specific pretraining on base models, and task-specific fine-tuning approaches.

The first factor is the amount of context required for accurate detection. Since each user may generate a large number of messages, the input size becomes a critical consideration. One of our goals was to investigate the impact of contextual information on system performance—that is, to assess how well different models perform depending on the amount of context they can handle. To this end, we evaluated three distinct systems: one based on classical machine learning approaches, another using a RoBERTa model, and a third employing a Longformer model. Each of these systems can process different input lengths, allowing us to analyze the effect of context size on prediction accuracy.

1. Classical machine learning approaches have no limit on the input size.
2. The selected RoBERTa model has a limit of 512 tokens in the input.
3. The selected Longformer model has a limit of 4096 tokens in the input.

The second factor our team wanted to investigate was the impact of domain-specific pre-training on base models and how it affects their ability to generate domain-relevant embeddings. This investigation was carried out by comparing two types of models:

1. The first group includes base models such as BERT and RoBERTa, which were pre-trained on large, general-domain corpora.
2. The second group consists of models built upon these base architectures, but further pre-trained on domain-specific data related to mental health.

The third factor is to explore the impact of task-specific fine-tuning compared to more general, classic fine-tuning approaches. For this purpose, we created two different datasets to train and evaluate the performance of the Transformer-based systems.

1. **Dataset 1.** We created only one sample per user by accumulating all their messages for high-risk and low-risk labeled users.
2. **Datasets 2.** If we had prior knowledge of the point at which a user begins to exhibit symptoms indicating a risk of mental illness, we could label all earlier messages as low-risk and the message containing the onset of symptoms and all subsequent messages as high-risk. This approach would allow us to increase the number of high-risk samples, potentially leading to a more accurate model. The data augmentation process used to implement this idea is described below. Using this technique, we obtain a dataset for each input length, resulting in two datasets in total (512 and 4096 tokens).

To carry out our experimentation, we divided the original dataset into training (80% of users) and evaluation (20% of users), maintaining the proportions of positive (high-risk users) and negative (low-risk users) samples in each partition. Table 3 shows the distribution of samples in Dataset 1 for Task 1.

	Training	Evaluation
Low risk	145	37
High risk	140	35
Total	285	72

Table 3
Distribution of samples in Dataset Task 1 for training and evaluation partitions.

Following the same procedure as in Task 1, we divided the corpus into two partitions: training (80%) and evaluation (20%), while preserving the class distribution in each partition. Table 4 shows the distribution of classes across both partitions.

	Training	Evaluation
Betting	69	18
Online Gaming	85	21
Trading and Crypto	109	28
Loot Boxes	22	5
Total	285	72

Table 4
Distribution of samples in Dataset Task 2 for training and evaluation partitions.

3.1. Classical Machine Learning Classifier Approach

To evaluate the importance of context, we used a classical machine learning classifier capable of processing the full input context. One of the main limitations of Transformer-based models is their difficulty in handling long texts due to restrictions on input size. This constraint can negatively impact the classification performance, as the input may not capture the entire sample, potentially losing valuable information.

First, we compared several classical machine learning classifiers. For this purpose, we used the Scikit-learn library [8], which provides a wide range of tools to support our experimentation. All classifiers were used with their default parameters to ensure a fair comparison, and no data preprocessing or prior analysis was applied. For feature extraction, we employed the TF-IDF method from the scikit-learn library, which generates a vector of the vocabulary size. The configuration used in this case corresponded to the same selected in the previous year's competition [9], which achieved the best performance in the experiments conducted. **Configuration:** "char_wb" , 4-5 n-gram

	Macro-P	Macro-R	Macro-F1
Linear SVM	**0.69**	**0.68**	**0.68**
Gradient Boosting	0.60	0.60	0.60
K-Neighboors	0.65	0.65	0.65
Random Forest	0.62	0.58	0.59

Table 5
The results from different classifiers in the evaluation partition in Task 1. The scores are the Macro-precision, recall, F1_score.

Table 5 shows the performance of the four classical approaches evaluated. The best-performing classifier was the Linear SVM on Precision and Recall, which is also reflected in a higher F1-score than the rest of the approaches. Since the training samples were the same for both tasks, we assumed that the SVM approach would obtain the best performance for Task 2; therefore, we chose this approach for both tasks. In addition to selecting the classification algorithm, we explored different preprocessing methods and incorporated additional information for each message:

- **Preprocess of Data:**
  1. *TweetTokenizer and stop words removal*: The text is tokenized using the TweetTokenizer, followed by removing stop words.
  2. *TweeTokenizer, cleaning and lematization*: This builds on the first approach by including additional preprocessing steps, such as cleaning the text, removing non-alphanumeric characters, and applying token lemmatization.
- **Sentimental Analysis:** We employed the Transformer-based model **"lxyuan/distilbert-base-multilingual-cased-sentiments-student"** [10] to perform sentiment analysis on each message per user. The model outputs three sentiment categories: positive, negative, and neutral. These outputs were normalized and incorporated as an additional feature along the TF-IDF representation.

To identify the optimal parameters for each model, we conducted an exhaustive search using GridSearch, a tool provided by Scikit-learn. The search was performed over the parameters `C`, `tol`, and `loss`. The possible values are:

- **C**: [$1e^{-1}$, 1, 10, 100]
- **tol**: [0.1, 0.01, 0.001, 0.0001]
- **loss**: [hinge, squared_hinge]

Task 1

In total, we obtained four different configurations for experimentation. Table 6 presents these configurations, with the optimal parameters for each one listed in the *Best Parameters* column.

	Data preprocess	Sentiment analysis	Best parameters
SVM-t1-1	1	No	'C': 10, 'loss': 'squared_hinge', 'tol': 0.1
SVM-t1-2	2	No	'C': 1, 'loss': 'squared_hinge', 'tol': 0.01
SVM-t1-3	1	Yes	'C': 1, 'loss': 'squared_hinge', 'tol': 0.01
SVM-t1-4	2	Yes	'C': 1, 'loss': 'hinge', 'tol': 0.1

Table 6
Summary of the different configurations of the SVM classifiers.

Table 7 shows the results on the evaluation partition. The best configuration was **SVM-t1-4**, corresponding to a Linear SVM combined with sentiment analysis and the second preprocessing approach that includes thorough data cleaning and preprocessing.

	Macro-P	Macro-R	Macro-F1
SVM-t1-1	0.68	0.68	0.68
SVM-t1-2	0.68	0.68	0.68
SVM-t1-3	0.70	0.70	0.69
SVM-t1-4	**0.71**	**0.71**	**0.71**

Table 7
Results of the different configurations of the SVM classifiers on the evaluation partition for Task 1. In bold, the best result for each metric.

Task 2

For Task 2, we performed the same hyperparameter optimization as in Task 1. The best values for the hyperparameters for each combination are shown in Table 8.

	Data preprocess	Sentiment analysis	Best parameters
SVM-t2-1	1	No	'C': 10, 'loss': 'hinge', 'tol': 0.1
SVM-t2-2	2	No	'C': 10, 'loss': 'hinge', 'tol': 0.01
SVM-t2-3	1	Yes	'C': 10, 'loss': 'hinge', 'tol': 0.1
SVM-t2-4	2	Yes	'C': 10, 'loss': 'hinge', 'tol': 0.01

Table 8
Summary of the different configurations of the SVM classifiers for Task 2.

Table 9 shows the results obtained by each combination for Task 2 in evaluation. It can be noticed that all the combinations achieved a perfect score (*F1-score* of 1). However, considering that all systems achieved a perfect score, we take into account the preprocessing cost. In this regard, we assume **SVM-t2-1** as a more favorable system due to its lower computational and preprocessing demands, as it does not require sentiment analysis or extensive data preprocessing.

	Macro-P	Macro-R	Macro-F1
SVM-t2-1	1.00	1.00	1.00
SVM-t2-2	1.00	1.00	1.00
SVM-t2-3	1.00	1.00	1.00
SVM-t2-4	1.00	1.00	1.00

Table 9
Results of the different configurations of the SVM classifiers on the evaluation partition for Task 2. In bold, the best result for each metric.

3.2. Straighforward Fine-tuning Approach

It is well-known that the Transformer architecture is the state-of-the-art of language models. In this shared task, we used two Transformer-based architectures: RoBERTa and Longformer [5, 6].

- **RoBERTa**: RoBERTa generally offers strong versatility and performance for classification tasks. However, these types of models present the inability to process input longer sequences; usually, a maximum of 512 tokens. This limitation poses a challenge for tasks involving long contexts, such as those addressed in this work. Therefore, we used a RoBERTa as a baseline to compare it against other models that can handle longer input sequences.
- **Longformer**: Longformer, short for "Long-Document Transformer," was specifically designed to efficiently process extended input sequences, making it more suitable than standard Transformer models such as BERT or RoBERTa for tasks involving long contexts. This architecture presents the following key features:
  - **New attention mechanism**: Longformer uses a sliding window attention mechanism, where each token attends only to a fixed number of neighboring tokens. This significantly reduces computational complexity compared to full self-attention.
  - **Global attention mechanism**: The model allows specific tokens to receive global attention, enabling them to attend to all other tokens in the sequence, while the rest remain in a local attention scope. This hybrid approach balances efficiency and contextual understanding.

One of the objectives of this model was to compare base models trained on general-domain data with models that had undergone additional training specifically in the mental health domain. To this end, we selected the following models:

1. **General-domain models:** We selected the models *PlanTL-GOB-ES/RoBERTa-large-bne* and *PlanTL-GOB-ES/Longformer-base-4096-bne-es* [11], developed by the Spanish government. The RoBERTa model is based on the original RoBERTa architecture and was pre-trained on the largest Spanish-language corpus, composed of texts from the National Library of Spain. The Longformer model adapts this RoBERTa model to the Longformer architecture, enabling the processing of longer input sequences. These models can be found in the Hugging Face repository [12].
2. **Specific-domain models**: Our team developed three specific models; two based on Longformer architectures and one built upon the RoBERTa architecture. These models were trained to generate contextual embeddings tailored to the mental health domain. For training, we used the Suicidal and Mental Health (SWMH) corpus [13], which contains texts related to a range of mental health disorders. We did not train the models from scratch; instead, we continued their pretraining to adapt them specifically to the mental health domain, as recommended by other research [14].

For each base model, we performed straightforward fine-tuning on each of them using **Dataset 1**. In this fine-tuning, we adjusted the models with the full context available for each user. This experimentation served as a baseline for comparing a more general approach with a more task-specific one. Table 10 shows the configuration used in the fine-tuning process.

Parameter	**Value**
optimizer	AdamW
learning rate	3e-5
lr scheduler type	linear
weight decay	0.01
number of epochs	10
training batch size	16

Table 10
Parameters for the fine-tuning process.

Task 1

Table 11 presents the results of the different models on the evaluation partition.

- **RoBERTa-t1-g**: Modelo PlanTL-GOB-ES/RoBERTa-large-bne
- **RoBERTa-t1-s**: RoBERTa model pre-trained on domain-specific mental health data.
- **Longformer-t1-g**: Modelo PlanTL-GOB-ES/Longformer-base-4096-bne-es
- **Longformer-t1-s**: Longformer model pre-trained on domain-specific mental health data.

The Longformer model, which was pre-trained with domain-specific mental health data, achieved the best results due to its ability to handle long texts and specialized pre-training. The results show that pre-training on domain-relevant data improved the model's understanding of specialized language and enhanced its adaptation to the specific tasks.

	Macro-P	**Macro-R**	**Macro-F1**
RoBERTa-t1-g	0.592	0.601	0.587
RoBERTa-t1-s	0.638	0.639	0.638
Longformer-t1-g	0.638	0.633	0.620
Longformer-t1-s	**0.652**	**0.650**	**0.649**

Table 11
RoBERTa's and Longformer's results for Task 1 on evaluation partition.

Task 2

We used the same models as in Task 1, but using a different dataset. In this case, all the models obtained the highest score, i.e., correctly predicting all the samples. The results obtained by the models on the evaluation partition are shown in Table 12.

	Macro-P	Macro-R	Macro-F1
RoBERTa-t2-g	1.00	1.00	1.00
RoBERTa-t2-s	1.00	1.00	1.00
Longformer-t2-g	1.00	1.00	1.00
Longformer-t2-s	1.00	1.00	1.00

Table 12
RoBERTa's and Longformer's results for Task 1 on evaluation partition.

3.3. Task Adaptive Fine-tuning

The third objective of our work was to investigate the impact of task-specific training, specifically tailored to an early detection scenario. To achieve this, we applied a data augmentation process to adapt the original dataset to the requirements of early detection. This process followed the same methodology employed in previous editions [9].

The data augmentation strategy aimed to generate additional samples for each positive user. This process involved identifying the specific message in which a user began to exhibit signs of a mental health disorder. Once the critical message was determined, each possible concatenation of the user's message history was labeled as low-risk if it occurred before the selected message, or high-risk otherwise.

To detect the critical message, we conducted a series of experiments in which we trained and evaluated various models to detect this critical message.

3.3.1. Best model for data augmentation

The experimentation phase aimed to identify the model with the highest performance in detecting the critical message. To this end, we replicated the inference procedure proposed by the competition: each model received input batches containing one message per user. The models were then required to predict the correct labels as early as possible following this procedure. For the experimentation, we used the complete dataset to identify the critical message for each positive user.

The models proposed for this analysis were **Longformer-t1-s**, **RoBERTa-t1-s**, and **SVM-t1-4**, as they achieved the best performance during the training phase. Table 13 presents the experimentation results, where we can observe that **Longformer-t1-s** achieved the best performance in classifying users using an early detection strategy, so this was the selected model.

	Macro-P	Macro-R	Macro-F1
SVM-t1-4	0.77	0.69	0.70
RoBERTa-t1-s	0.85	0.77	0.78
Longformer-t1-s	**0.88**	**0.79**	**0.81**

Table 13
RoBERTa's and Longformer's results for Task 1 on evaluation partition.

This technique results in a new dataset with a higher number of positive samples for training. Table 14 shows the two datasets created using the data augmentation process described above, one for each input length: 512 tokens for RoBERTa and 4096 tokens for Longformer. **Dataset 2** refers to the one created with a maximum message length of 512, while **Dataset 3** refers to the one with a maximum length of 4096.

	Low Risk	High risk
Original	145	140
Dataset 2	6924	6412
Dataset 3	9766	8582

Table 14
Distribution of samples in the new dataset: train partition.

Task 1

Table 15 presents the results of the models with a straightforward fine-tuning process (RoBERTa-t1-s and Longformer-t1-s) and the ones obtained with data augmentation (RoBERTa-t1-t and Longformer-t1-t). As observed, the models fine-tuned with the task-adaptive approach achieved better evaluation performance, indicating that a fine-tuning process that emulates the early detection aspect could align better with the specific task, leading to improved results.

	Macro-P	Macro-R	Macro-F1
RoBERTa-t1-s	0.638	0.639	0.638
RoBERTa-t1-t	0.692	0.695	0.691
Longformer-t1-s	0.652	0.650	0.649
Longformer-t1-t	**0.767**	**0.757**	**0.759**

Table 15
RoBERTa's and Longformer's results for Task 1 on evaluation partition.

Task 2

For this task, we did not submit any system using this approach for several reasons:

1. The first reason was that this task does not involve early detection; therefore, adapting our models to label the users as early as possible is unnecessary. Instead, we decided that the optimal strategy was to wait until the maximum amount of context was available before making an inference, corresponding to fine-tuning using all the samples per user.
2. The second reason was related to the nature of the task itself. Since all users exhibit some form of addiction, there was no critical message that marks a transition from a non-addicted to an addicted state.
3. The third reason was that the models previously tested had already demonstrated strong performance on this task. Consequently, and considering time constraints, we opted to limit further experimentation.

4. Runs

Table 16 summarizes the selected model for each run, along with the corresponding performance on the evaluation set.

The rationale behind selecting these models is to validate the hypotheses of our research: to assess the impact of context on prediction performance, to evaluate how different models handle long-context inputs, to compare models specifically pre-trained in the mental health domain against those that are not (such as the competition baseline), and finally, to explore the effect of task-specific fine-tuning for model adaptation.

4.1. Run Configuration

In addition to selecting the model for each run, the classification systems required configuring additional parameters.

	Task	Model	Macro-P	Macro-R	Macro-F1
Run0	1	SVM-t1-4	0.710	0.710	0.710
Run1	1	Longformer-t1-t	**0.767**	**0.757**	**0.759**
Run2	1	RoBERTa-t1-t	0.692	0.695	0.691
Run0	2	SVM-t2-1	1.000	1.000	1.000
Run1	2	Longformer-t2-s	1.000	1.000	1.000
Run2	2	RoBERTa-t2-s	1.000	1.000	1.000

Table 16
Summary of the approaches chosen for each run. Also, the performance achieved by each system in the evaluation partition is considered.

Task1:

- For each round of the competition, the input to the classifier was a newly generated sample composed of the user's latest message concatenated with all previous messages.
- Each system employed an initial context threshold; predictions were deferred until the accumulated context reached a minimum of 100 tokens.
- Both the RoBERTa and Longformer systems have a maximum token limit. Once this limit was reached, the system returned the most recent prediction without further updates.

Task2:

The second task only considers the most recent message sent by each user. However, since we could not predict when a user would stop receiving new information, we chose to send new predictions at each round.

5. Results

5.1. Task 1

Table 17 presents the results achieved by our team in Task 1. The structure of Table 17 is as follows: each row corresponds to a run, with a special row highlighting the highest values in the competition. The systems in the competition were ranked based on the Macro-F1 score (last column).

	Model	Accuracy	Macro-P	Macro-R	Macro-F1
Run0	SVM	0.538	0.535	0.535	0.533 (6)
Run1	Longformer	**0.550**	**0.564**	**0.556**	**0.540 (4)**
Run2	RoBERTa	0.538	0.543	0.542	0.534 (5)
Highest	-	0.569	0.568	0.567	0.567 (1)

Table 17
Results for the 3 runs on Task 1. *Highest* refers to the highest values achieved in the competition. The values inside the parentheses indicate our position in the ranking.

Table 17 presents the results, highlighting **Run 1** corresponding to the **Longformer-t1-t**, a specifically pre-trained Longformer model fine-tuned with data augmentation, as the best-performing system. This run secured fourth place in the competition. In light of the results, the following conclusions have been reached:

- The Longformer achieved the best performance among our runs, demonstrating its effectiveness in handling long contexts more efficiently. This confirms that tasks involving numerous messages or lengthy texts deliver superior results.
- When comparing our best run to the baseline, which achieved a Macro-F1 score of *0.428*, we observe a superior performance from our system. The baseline is based on RoBERTuito [15], a RoBERTa-based model pre-trained on Spanish social media text. This result supports our two

hypotheses: task-specific pretraining and tailored fine-tuning contribute to improved model performance.

- Although the top-performing runs employed Transformer-based models, the SVM run achieved a comparable result, only 0.1% lower than *Run2* and 1% less than *Run1*. This suggests that classical approaches such as SVMs remain effective for detecting mental health conditions, owing to their capacity to manage extensive contextual information. Consequently, SVMs are a suitable option in scenarios with limited computational resources.

5.2. Task 2

Table 18 shows the results for Task 2. We achieved third place with our Run1.

	Model	Accuracy	Macro-P	Macro-R	Macro-F1
Run0	SVM	0.881	0.895	0.829	0.832 (8)
Run1	Longformer	**0.913**	**0.929**	**0.877**	**0.887 (3)**
Run2	RoBERTa	0.888	0.904	0.862	0.873 (4)
Highest	-	0.938	0.952	0.915	0.927 (1)

Table 18
Results for the three runs on Task 2. *Highest* refers to the highest values achieved in the competition. The values inside the parentheses indicate our position in the ranking.

As shown in Table 18, the results obtained by all runs are highly competitive, although not as strong as with the development split. These results further reinforce the previously discussed conclusions, supporting our initial hypotheses.

5.3. Carbon emission

One of the primary objectives of the competition is to identify systems capable of completing tasks with minimal resource consumption. This will assist in identifying technologies that can operate on mobile devices or personal computers, as well as those with the lowest carbon emissions. Consequently, we provide the following information:

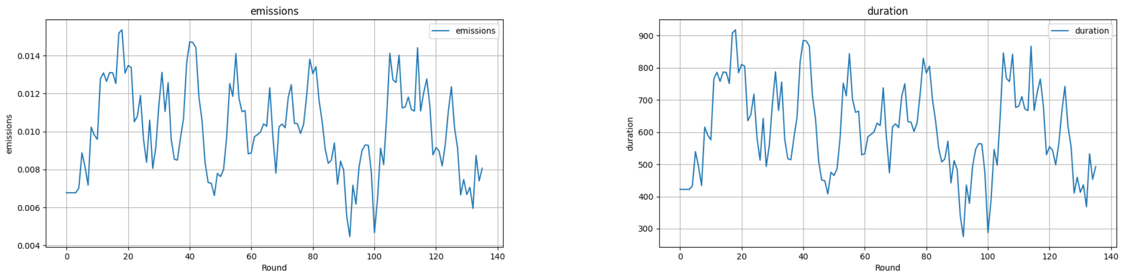
- Total time to process (in milliseconds)
- Kg in CO_2 emissions.

Using the provided script, which leverages the CodeCarbon API [16] to calculate emissions, we present our team's computer configuration in Table 19. This table outlines the types and quantities of CPUs and GPUs utilized and the total amount of RAM employed. The results for the **Longformer-t1-t Run 1** are also presented.

Measurements	Values
CPU_Count	24
GPU_Count	1
CPU_Model	12th Gen Intel(R) Core(TM) i9-12900K
GPU_Model	NVIDIA GeForce RTX 4090
RAM_Total_Size	128 GB
Country_ISO_Code	ESP

Table 19
Computer configuration

Figure 1 presents the variation in emissions and duration observed throughout the experimentation process. A clear correlation between these metrics indicates that rounds of longer duration resulted in higher CO_2 emissions. Given that all rounds employed identical models and configurations, the



(a) Emissions of CO_2 (Kg) of each round

(b) Duration (milliseconds) of each round

Figure 1: Emissions and Durations Graphs

primary factors influencing emissions were the duration of each round and the cumulative context associated with the user.

Figure 2 illustrates the cumulative energy consumption of each component. The GPU emerges as the predominant energy consumer, representing approximately 96% of total energy usage. The RAM follows with a consumption of 2.5%, while the CPU contributes only 0.2% to the overall energy consumption.

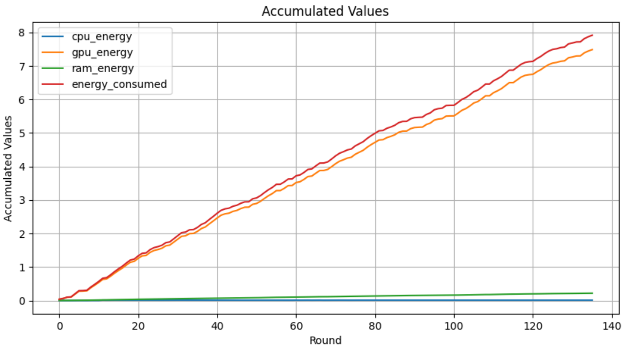


Figure 2: Accumulated values of energy (kWh) during the rounds

6. Task 2 Analysis

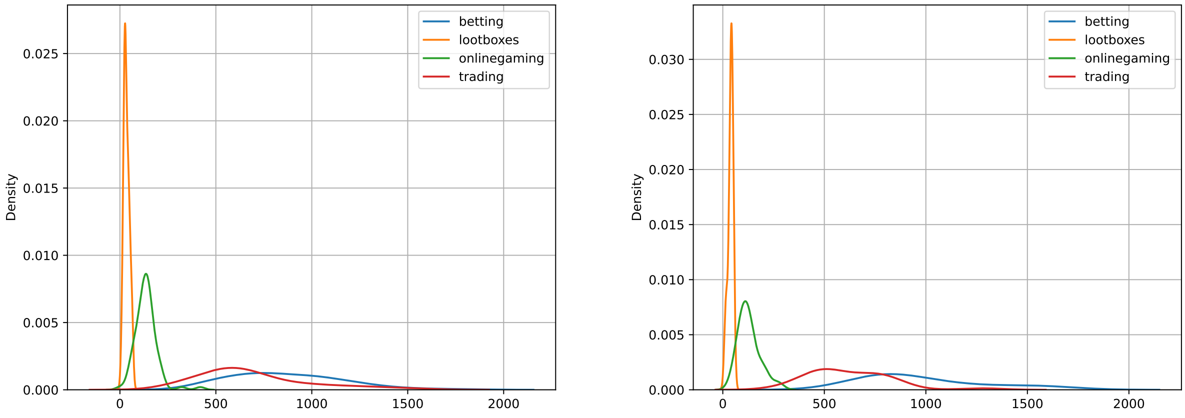
After analyzing the results obtained in Task 2, we identified a significant discrepancy between the evaluation and test results. Moreover, all models achieved perfect scores during the evaluation phase, which raised further concerns. These observations prompted us to examine the data provided for this task more closely and explore potential solutions.

Our first step was to verify the Train and Evaluation splits for potential errors. However, we consistently obtained the same results after checking the class distributions in each partition and conducting training with different splits. Having ruled out any issues with the data partitioning, we proceeded to analyze the provided dataset in greater detail. Specifically, we conducted a class-by-class analysis to identify any possible anomalies or patterns. The study focused on examining the message length, in terms of tokens, for each class.

Figure 3 shows the density curves for each class in each data partition. We observe a notable disparity between classes: two contain significantly shorter texts than the others. If we sort the classes by text length, the resulting ranking from shortest to longest is as follows: (1) *Lootboxes*, (2) *Online Gaming*, (3) *Trading and Crypto*, (4) *Betting*. When examining each class individually, we observe that the *Lootboxes* class exhibits a significantly higher density than the others, with most of its texts concentrated around 30 tokens. Subsequently, the *Online Gaming* class shows a lower density; however, most of its texts still fall within the range of 100 to 150 tokens. Lastly, the last two classes display similar density curves, with no predominant text length. Instead, the texts are distributed over a broad range, approximately from 300 to 1500 tokens.

When focusing on partition-based analysis, we can observe that previously described patterns are consistent across partitions. Although there is a slight variation between the *Trading and Crypto* and *Betting* classes, where the density curves are not as similar as before, the overall trend remains: the curves in both partitions are broadly comparable.

For additional statistics related to the class analysis, please refer to Appendix A, specifically Table 20.



(a) Train

(b) Evaluation

Figure 3: Distribution of text lengths across the four classes and each data partition. The X-axis represents the number of words, while the Y-axis shows the density.

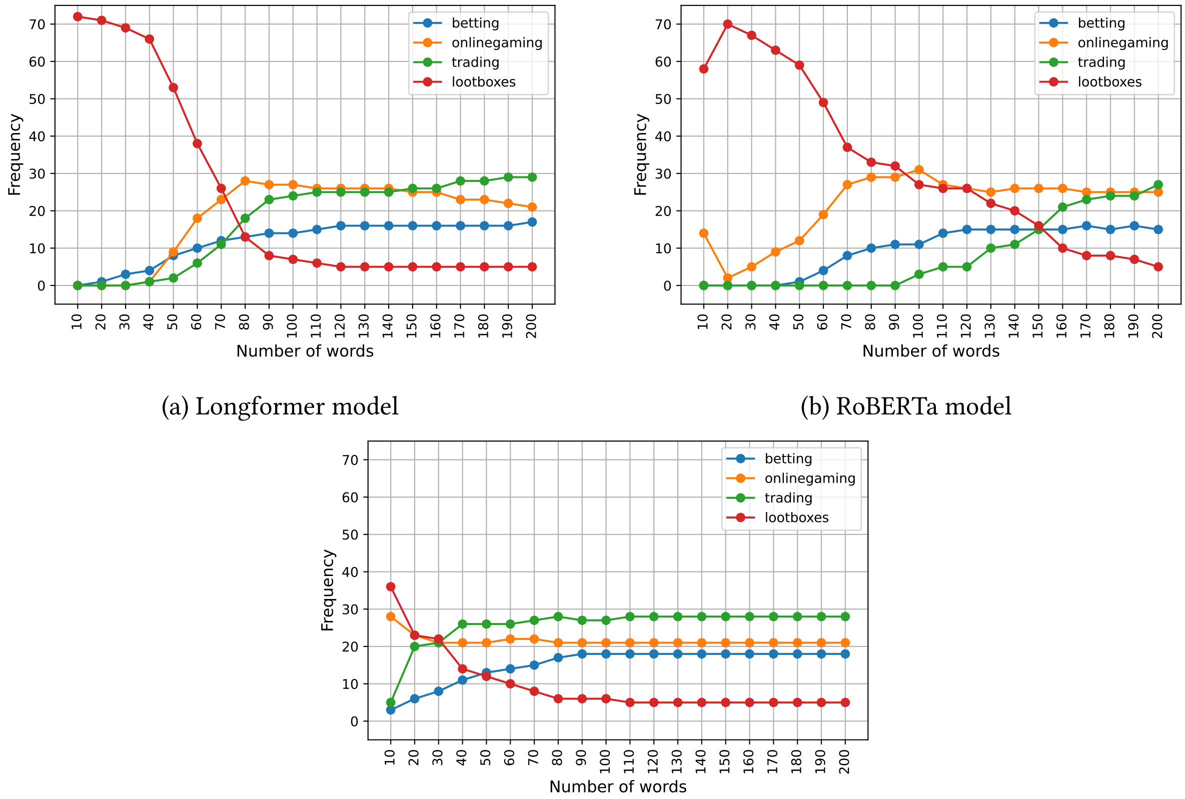
All these observations have led us to hypothesize that the different models may be leveraging message length as a feature when classifying samples during inference. We are not suggesting that it is the primary factor driving classification decisions, but rather that it is a highly influential one. This is especially relevant given that there is a clear distinction in message lengths across classes, and this pattern is also preserved in both the training and evaluation partitions. As a result, models may inadvertently rely on this feature during training, as it is not penalized; on the contrary, it may even be reinforced.

To test this hypothesis, we conducted the following experiment: we simulated the competition setup where, in each round, one message per user was provided. However, we supplied only 10 words per user instead of the whole message. This allowed us to evaluate the models' predictions under conditions of reduced text length and assess the extent to which their performance depends on message length.

Figure 4 shows the distribution of predictions by class and by each model submitted to the task. We observe that models based on Transformer architectures frequently predict the lootboxes class when provided with messages containing very little contextual information, even though this class presents the smallest number of samples. Coincidentally, *lootboxes* is also the class with the shortest messages overall. Furthermore, as the size of the available context increases (number of words already received), the prediction frequencies become more balanced, leading to improved performance.

In contrast, the SVM model is less sensitive to message length and relies more heavily on the vocabulary. This is evident because it does not initially overpredict lootboxes, but instead shows a more diverse distribution of predictions from the start. Nevertheless, we still observe that increasing the number of tokens leads to a more balanced prediction distribution and, consequently, better performance. This could be related to the TF-IDF extraction feature, which was fitted with all the user messages treated as a single document (the messages were concatenated).

The experiment supports our hypothesis that the models have leveraged this unintended feature for classification. This reliance can lead the models to make errors during testing if the test partition does not follow the same distribution as the training and evaluation partitions.



(a) Longformer model

(b) RoBERTa model

(c) SVM model

Figure 4: Evolution of the prediction of the different models. The X-axis represents the number of context words used in each prediction, while the Y-axis shows the number of labels of each class.

7. Conclusion

In this paper, we have presented the participation of the ELiRF-VRAIN team in the shared tasks of MentalRiskES at IberLef 2025. Besides evaluating traditional classification models and cutting-edge Transformer models, our team's most innovative contribution was the use of Longformer models to broaden the context for decision-making, leveraging pre-trained models specifically designed for the mental health domain, and introducing a new data augmentation technique that customizes model training to the specific task at hand.

The highly competitive results support our proposal's validity, demonstrating the performance of models tailored explicitly for the task at hand.

For future work, two areas of improvement are identified. Firstly, we aim to enhance early detection so that the system requires less initial context to make accurate decisions. Secondly, we plan to incorporate Explainable Artificial Intelligence (XAI) techniques to understand the system's behavior better. Lastly, we aim to introduce new analyses and explore alternative training techniques to avoid that systems take into account the length of the input as a classification feature.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2021-126061OB-C41. Partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València PAID-01-23. It is also partially supported by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training and by the Generalitat Valenciana under CIPROM/2021/023 project.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Grammarly in order to: Grammar and spelling check, Paraphrase, translate, and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] World Health Organization, Mental disorders, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders, accessed: 2024-05-15.

[2] 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).

[3] 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.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All You Need, Advances in Neural Information Processing Systems 30 (2017). URL: https://arxiv.org/abs/1706.03762, accessed: 2024-05-15.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[6] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The Long-Document Transformer, arXiv preprint arXiv:2004.05150 (2020). URL: https://arxiv.org/abs/2004.05150.

[7] 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.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

[9] A. Casamayor, V. Ahuir, A. Molina, L.-F. Hurtado, ELiRF-VRAIN at MentalRiskES 2024: Using LongFormer for Early Detection of Mental Disorders Risk, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024) co-located with 40th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), volume 3756 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 24–33. URL: https://ceur-ws.org/Vol-3756/MentalRiskES2024_paper3.pdf.

[10] L. X. Yuan, distilbert-base-multilingual-cased-sentiments-student (revision 2e33845), 2023. URL: https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student. doi:10.57967/hf/1422.

[11] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, MarIA: Spanish Language Models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, 2020. URL: https://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[13] S. Ji, X. Li, Z. Huang, E. Cambria, Suicidal ideation and mental disorder detection with attentive relation networks, Neural Computing and Applications (2021).

[14] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, CoRR abs/2004.10964 (2020). URL: https://arxiv.org/abs/2004.10964. arXiv:2004.10964.

[15] J. Cañete, G. Chaperon, C. Fuentes, J. Pérez, B. Poblete, RoBERTuito: A pre-trained language model for social media text in Spanish, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT), Association for Computational Linguistics, 2022, pp. 132–140. URL: https://aclanthology.org/2022.wnut-1.14.

[16] CodeCarbon, CodeCarbon: Track and Reduce Your Carbon Emissions from Machine Learning Workloads, https://mlco2.github.io/codecarbon/index.html, 2024. Accessed: 2024-05-15.

A. Statistics of Number of Total Number of Words

Table 20 presents the extended statistics of the different classes across the various partitions. For each class, we report the number of users/samples, the mean, the standard deviation, the median, the first and third quartiles, and the minimum and maximum values for the total number of words per text.

Partition	Label	N	Mean	Std	Median	Q1	Q3	Min	Max
Train	Lootboxes	21	34.62	13.74	29.0	25.0	44.0	12	63
Train	Online Gaming	85	135.80	55.99	135.0	108.0	158.0	1	419
Train	Betting	70	867.66	288.96	816.5	645.25	1030.75	375	1782
Train	Trading	109	693.43	292.36	615.0	516.0	812.0	187	1579
Eva	Lootboxes	6	38.17	10.96	40.5	36.5	45.25	16	50
Eva	Online Gaming	21	128.38	51.07	117.0	100.0	143.0	53	271
Eva	Betting	17	983.71	291.44	867.0	753.0	1142.0	619	1638
Eva	Trading	28	622.29	199.58	554.5	481.75	749.25	277	1277

Table 20
Descriptive statistics of text lengths for each class in the Train and Evaluation partition.