Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 435-448 recibido 08-07-2024 revisado 15-07-2024 aceptado 30-07-2024

Overview of MentalRiskES at IberLEF 2024: Early Detection of Mental Disorders Risk in Spanish

Resumen de la Tarea MentalRiskES en IberLEF 2024: Detección Precoz del Riesgo de Trastornos Mentales en Español

Alba María Mármol-Romero,[1] **Adrián Moreno-Muñoz,**[1] **Flor Miriam Plaza-del-Arco,**[2]
M. Dolores Molina-González,[1] **M. Teresa Martín-Valdivia,**[1]
L. Alfonso Ureña-López,[1] **Arturo Montejo-Ráez**[1]
[1]University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]Bocconi University, Via Sarfatti 25, 20100, Milan, Italy
[1]{amarmol, ammunoz, mdmolina, maite, laurena, amontejo}@ujaen.es,
[2]{flor.plaza@unibocconi.it}

Abstract: This paper presents the MentalRiskES shared task organized at IberLEF 2024, as part of the 40th International Conference of the Spanish Society for Natural Language Processing. This task aims to promote the early detection of mental risk disorders in Spanish. We propose three detection tasks: Task 1 to detect risk for depression or anxiety, Task 2 to detect risk for depression or anxiety but determining contextual risk factors and Task 3 to identify whether a subject is at risk for suicidal ideation. Furthermore, we asked participants to submit measurements of carbon emissions for their systems, emphasizing the need for sustainable natural language processing practices. In this second edition, 28 teams registered, 12 submitted results, and 10 presented papers. Most teams experimented with Transformers, including features, data augmentation, and preprocessing techniques.
Keywords: mental disorder risk detection, early detection of anxiety, early detection of depression, early detection of eating disorders.

Resumen: Este artículo presenta la tarea MentalRiskES en IberLEF 2024, como parte de la 40ª edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. El objetivo de esta competición es promover la detección temprana de trastornos mentales en español. Proponemos tres tareas de detección precoz: Tarea 1 para detección de riesgo de depressión o ansiedad, Tarea 2 para detección de riesgo de depressión o ansiedad pero determinando los factores contextuales de riesgo y Tarea 3 para identificar si un sujeto tiene riesgo de sufrir de ideación suicida. Además, pedimos a los participantes que enviaran mediciones de las emisiones de carbono de sus sistemas, haciendo hincapié en la necesidad de prácticas sostenibles de procesamiento del lenguaje natural. En esta segunda edición, 28 equipos se registraron, 12 enviaron predicciones y 10 presentaron artículos. La mayoría experimentó con Transformers, incluyendo características, ampliando datos y técnicas de preprocesamiento.
Palabras clave: detección precoz de trastornos mentales, detección precoz de ansiedad, detección precoz de depresión, detección precoz de trastornos alimentarios.

1 Introduction

Technology is constantly evolving and more activities can be performed on screen-based devices. With this growing participation in the digital world, social networks have become increasingly popular, especially among younger people (Andreassen, Pallesen, and Griffiths, 2017). Numerous evidence suggests a notable connection between young people's excessive participation in social networks and severe adverse mental health consequences, namely increased symptoms of depression and anxiety, as well as higher levels of stress (Shannon et al., 2022). On the other hand, according to the World Health Organi-

ISSN 1135-5948 DOI 10.26342/2024-73-33

sation (WHO),[1] each year, more people die as a result of suicide than HIV, malaria or breast cancer or war and homicide. Moreover, the link between suicide and mental illness (especially depression) is well-established in high-income countries, but many suicides occur impulsively in times of crisis. The COVID-19 pandemic caused a huge crisis that remains today, and many risk factors such as job loss, financial stress and social isolation are once again very present. Therefore, the attention to detect mental health problems and suicide prevention are now more important and necessary than ever.

In the last few years, to detect mental health problems such as depression, anxiety or suicidal ideation from user-generated textual data, researchers have increasingly turned to Natural Language Processing (NLP) and deep learning. In fact, relevant evaluation campaigns such as the Cross-Lingual Evaluation Forum (CLEF) have hosted over the last few years the Early-Risk Identification (eRisk) task to address the detection of symptoms of depression (Parapar et al., 2023) or Early Detection of Signs of Self-Harm (Parapar et al., 2021) among other disorders. These tasks have been mainly focused on texts written in English, leaving aside other of the most widely used languages in the world, such as Spanish.

This paper describes the second edition of a novel task on early risk identification of mental disorders in Spanish comments from social media sources. The first edition (Mármol-Romero et al., 2023) took place last year in the Iberian Languages Evaluation Forum (IberLEF) as part of the International Conference of the Spanish Society for Natural Language Processing (SEPLN) 2023. The task was resolved as an online problem, that is, the participants had to detect a potential risk as early as possible in a continuous stream of data. Therefore, the performance not only depends on the accuracy of the systems but also on how fast the problem is detected. These dynamics are reflected in the design of the tasks and the metrics used to evaluate participants. For this second edition, we propose three novel tasks, the first subtask is about the detection of the disorder, the second subtask consists of detecting the context that may be associated with the disorder, and the third subtask is about suicidal ideation detection.

[1] https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide

2 Tasks

In this section, we describe the different tasks proposed in the second edition of the competition.

2.1 Task 1: Disorder detection

Detect if a user suffers from depression or anxiety, or if there is no detected disorder at all. This is a multiclass task with three possible labels: depression, anxiety or none.

- **depression**: depression is characterised by persistent sadness, low mood, and a lack of interest or pleasure in activities that were previously rewarding and pleasurable. A user is considered to be suffering from depression when he/she expresses everyday situations, desires, or actions related to the suffering of such pathology
- **anxiety**: anxiety disorders are recognized by feeling intense, excessive and persistent worries, restlessness and fears about daily situations. A user is considered to be suffering from the disorder when he/she expresses everyday situations, desires or actions related to the suffering of such pathology.
- **none**: the user does not present evidence of suffering from any of the before disorders.

It was important to note that there may be users who present in their messages symptoms of anxiety and depression, however, there is one more predominant than the other which is the one to be indicated.

2.2 Task 2: Context detection

The first part of this task consists of detecting if the user suffers from depression or anxiety, or if there is no detected disorder at all. This is the same as the task 1 described before. In addition to detecting the disorder, participants must detect the context where the problem seems to come from. Available contexts are:

- **addiction**: disorder is influenced by the presence of an addiction disorder, such as substance use, pathological gambling, and alcoholism, among others.

- **emergency**: disorder is influenced by exceptional external factors such as pandemics, war conflicts, and natural disasters, among others.
- **family**: disorder is influenced by family problems.
- **work**: disorder is influenced by work-related problems.
- **social**: disorder is influenced by social problems.
- **other**: it is detected a context that is not among the previous ones.
- **none**: it is no detected a context.

2.3 Task 3: Suicidal ideation detection

Detect if a user is manifesting symptoms of potential suicidal ideation. Labels are 0 for "control" (negative, the user does not suffer from potential suicidal ideation) or 1 for "suffer" (positive). Only test data was provided for this task.

- **suffer**: suicidal ideation is characterized by a thought about not wanting to want to live or not planning to take one's own life. A user is considered to be suffering from the pathology when he/she has knowledge of the subject and applies it in his/her daily life: he/she expresses daily situations, wishes or actions related to suffering from the pathology.
- **none**: the user does not present evidence of suffering these symptoms.

2.4 Evaluation measures

Tasks are evaluated based on their specific definitions. We assess a system's performance using two primary criteria: **absolute classification** and **early detection effectiveness**. Table 1 showed the evaluation perspective for each task (each task needs a different way to be evaluated due to the nature of the decisions requested) and the metrics used to evaluate them.

2.4.1 Classification-based evaluation

This evaluation method focuses on binary or multi-class classification decisions made by the participating systems for each user. The objective is to determine whether a user is at risk of experiencing a mental health issue. To evaluate Tasks 1, 2, and 3, we used classical metrics such as accuracy, macro-precision, macro-recall, and macro-F1. These metrics assess the systems' final predictions after analyzing all posts from each subject in the dataset. For system ranking purposes, we selected the **macro-F1** metric as the primary criterion.

Tasks	Evaluation perspective	Metrics
1, 2, 3	Absolute multi-class and binary classification	Accuracy, Macro-P, Macro-R **Macro-F1**
2	Absolute multi-label classification	Accuracy, Macro-P, Macro-R **Macro-F1**, Micro-P Micro-R Micro-F1
1, 2, 3	Early detection in multi-class and binary classification	ERDE5, **ERDE30**, latencyTP, speed, latency-weightedF1

Table 1: Metrics used in evaluating submissions to MentalRiskES 2024 tasks. The reference metric (for submission ranking) for that evaluation is in bold.

2.4.2 Latency-based evaluation

We draw on the established framework from eRisk (Parapar et al., 2021) to derive metrics for measuring the early detection of positive subjects by the participating systems. For evaluating Tasks 1, 2 and 3, we used early risk evaluation metrics such as ERDE (Losada and Crestani, 2016) (ERDE5 and ERDE30), latencyTP, speed, and latency-weighted F1 (Sadeque, Xu, and Bethard, 2018). Given the short length of messages in our dataset, we determined that a larger number of messages is necessary for accurate early detection, leading us to prioritize the **ERDE30** metric for system ranking.

For early detection in multi-class classification (Tasks 1 and 2), we evaluate systems based on whether a user is classified as positive or not.

2.4.3 Efficiency metrics

Efficiency metrics are intended to measure the impact of the system in terms of resources needed and environmental issues. These metrics are not used to rank the system but to recognize those whose carbon footprint is environmentally friendly. So, we use metrics to measure the level of carbon emission produced for a system while it is predicting.

We aim to recognize systems capable of performing tasks with minimal resource demand. This allows us to identify technologies that can operate on mobile devices or personal computers and those with the lowest carbon footprint. To achieve this, each final prediction will include the following information:

- Minimum, maximum, mean, and variance of prediction time.
- Minimum, maximum, mean, and variance of CO2 emissions generated per prediction.
- Minimum, maximum, mean, and variance of energy consumption per CPU/GPU (kW) during prediction.
- Minimum, maximum, mean, and variance of RAM energy consumption (kW) during prediction.
- Minimum, maximum, mean, and variance of total energy consumption (kW) combining CPU, GPU, and RAM.
- Number and models of CPUs/GPUs used, the total RAM size required and 3-letter alphabet ISO Code of the respective country.

Participants used the CodeCarbon package[2] to track emissions, measured in kilograms of CO2-equivalents (CO2eq), to estimate the carbon footprint of their system predictions.

3 Dataset

This section describes the datasets used in each of the tasks.

3.1 Task 1 and task 2

In this edition, for tasks 1 and 2 of the shared task, we utilized threads of messages extracted from the MentalRiskES Corpus (Mármol-Romero et al., 2024). These messages were labelled as positive or negative indicators of depression or anxiety. The training and trial subsets of the data included all subjects from MentalRiskES2023 (Mármol-Romero et al., 2023) that use the data of the same corpus. For the test subset, we incorporated new subjects also contents from the MentalRiskES Corpus.

During the corpus creation process, special attention was given to contextual information, which had been annotated concurrently.

3.2 Task 3

The data used for task 3 were extracted following a process very similar to the one used for the creation of the MentalRiskES corpus. The data was obtained from Telegram.[3] Using Prolific[4] for annotator recruitment and Doccano (Nakayama et al., 2018) as the annotation platform. As in the previous edition, annotation guidelines were created. We used Cohen's kappa (Cohen, 1960) to measure the level of agreement between the annotators. We calculated it for each subset of data we released and took into account the level of agreement among 10 annotators. Cohen's kappa scores for the dataset used in task 3 are equal to 0.550 (moderate).

3.3 Dataset statistics

A total of three datasets are presented, covering depression, anxiety, and suicidal ideation. The first two datasets were combined to create a single dataset for use in tasks 1 and 2. Each dataset contains a collection of subjects with a list of messages they sent to a Telegram group. These subjects were split into 3 sets: (1) trial: to test the server, (2) train: to train systems, and (3) test: to test systems. In total, there are 885 subjects for tasks 1 and 2 and 55 subjects for task 3. The distribution of subjects in the sets and tasks can be seen in Table 2 and Table 3 shows a summary of the messages' distribution in each context and set.

The train and trial sets were sent to the participant as a .zip file containing JSON files. Each JSON contained a history of messages for a subject with the attributes: (1) id message, to identify the message; (2) message, the text message; and (3) date, the date and time when the message was sent to the group. On the other hand, to test the server, the trial set, again, and the test set was sent by the get request on a server whose response was a JSON file that contained a collection of messages from a lot of different subjects in one specific round. This process is repeated until all the messages from all the subjects are sent. The attributes for each JSON were:

[2]https://mlco2.github.io/codecarbon/index.html

[3]https://telegram.org/

[4]https://www.prolific.com/

	Task 1						Task 3			
	Anxiety		Depression		Control		Suicidal Ideation		Control	
	Subjs.	Msgs.	Subjs.	Msgs.	Subjs.	Msgs.	Subjs.	Msgs.	Subjs.	Msgs.
Trial	5	266	5	299	10	656	-	-	-	-
Train	88	14,265	164	12,909	213	19,237	-	-	-	-
Test	100	16,979	100	7,622	200	7,742	38	2,446	17	626

Table 2: Number of subjects and messages' distribution by set (trial, train and test) for each group of subjects used in each task.

		Trial	Train	Test
Addiction	Anxiety	0	622	340
	Depression	0	509	677
Emergency	Anxiety	0	1,364	1,696
	Depression	0	1,174	851
Family	Anxiety	0	4,410	5,697
	Depression	0	2,980	2,303
Work	Anxiety	0	3,099	4,186
	Depression	0	809	1,287
Social	Anxiety	51	3,912	6,161
	Depression	164	5,233	4,353
Other	Anxiety	176	3,593	6,237
	Depression	20	3,028	1,057
None	Anxiety	39	2,981	663
	Depression	115	3,783	1,267

Table 3: Number of messages' distribution by context (addiction, emergency, family, work, social, other, none) and by set (anxiety and depression).

(1) id message, to identify the message; (2) nick, to identify the subject; (3) round, to identify the round; (4) message, the text message; and (5) date, the date and time when the message was sent to the group.

4 Baselines

To establish a baseline benchmark for the tasks, we performed experiments using three different Transformer-based models. We experimented, as in the before edition of the shared task, with Spanish pre-trained models such as RoBERTa Base and RoBERTa Large, both from the MarIA project (Fandiño et al., 2022), and a multilingual pre-trained DeBERTa model (He et al., 2021). These models have demonstrated favourable results in Spanish tasks. In addition, RoBERTa Base,[5] RoBERTa Large[6] and mDeBERTa[7] are available at the HuggingFace models' hub.[8]

For experiments in tasks 1 and 2, we trained using the training set, used the trial set for early stopping, and evaluated using the test set. The experiments with Transformer used default hyper-parameters, however, we applied a fine-tuning that is specified in Table 4 and added a TrainerCallback to handle early stopping. All the training and evaluation experiments were performed on a node equipped with 2 NVIDIA RTX 4000 SFF Ada Generation.

Hyperparameters	Value
Learning Rate	5e-5
Weight Decay	0
Batch size	8
Seed	42
Max length	512
Number of train epochs	15
Early stopping patience	5

Table 4: Baselines training details for transformers-based experiments.

4.1 Task 1: Multiclass classification

In the HuggingFace transformer training arguments, the number of labels was set to 3 and the problem type was set to single-label

[5] `PlanTL-GOB-ES/RoBERTa-base-bne`
[6] `PlanTL-GOB-ES/RoBERTa-large-bne`
[7] `microsoft/mDeBERTa-v3-base`
[8] `https://huggingface.co`

classification. Early stopping was set to stop when the highest value in the macro-averaged F1 score was reached. It needed 12 epochs for DeBERTa, 6 for RoBERTa Large and 2 for RoBERTa Base.

4.2 Task 2: Two-level classification

For the first level, we use the same predictions calculated for task 1. For the second level, in the HuggingFace transformer training arguments, the number of labels was set to 7 and the problem type was set to multi-label classification. Early stopping was set to stop when the highest value in the macro-averaged F1 score was reached. It needed 8 epochs for DeBERTa, 14 for RoBERTa Large and 8 for RoBERTa Base.

4.3 Task 3: Binary classification

In this task, we use two simple baselines: one where all subjects were labelled positive and another where all subjects were labelled negative.

5 Participant approaches

- **Ixa-Med** (Larrayoz et al., 2024). This team participated in Task 1, using a message-level heuristic re-labelling approach based on the cosine similarity of embedding vectors to enhance consistency in supervised classification. They evaluated two language models, SBERT and BETO, and used a simple neural network to classify messages and determine the mental health status of users.
- **BUAP_01** (González and Vidal, 2024). This team participated in Task 1. They used a multilayer perceptron neural network and explored features such as sentiment information, publication dates and times, and probabilistic information. Three configurations of their model were tested, with Run 1 achieving the best results, using sentiment tagging combined with average nocturnal activity and probabilistic information.
- **UC3M-DAD** (Muñoz-Muñoz, Marco-Perez, and Ramirez, 2024). The approach followed strongly relies on sentiment analysis. The authors fine-tune RoBERTuito and BETO models, previously trained for sentiment analysis, on emotion detection (so only emotion-related texts are considered) and further fine-tune for disorder detection. Text is preprocessed removing stopwords.
- **Vteam** (Cedillo-Castelán, 2024). This team only participated in Task 3. They explore the detection of suicidal ideation using a range of machine-learning models integrating lexical features, encompassing both traditional algorithms (Logistic Regression, Naïve Bayes, Random Forest,..) and advanced Transformer models such as RoBERTuito or BERT.
- **UNED-GELP** (Fernandez-Hernandez et al., 2024). This team participated in Tasks 1 and 3. For Task 1 prediction, they used two different systems: one based on a two-step approach with BETO, and the other system relies on ANN techniques. For Task 3, they also employed two different systems: one based on BETO with training, and another based on a dictionary.
- **UMUTeam** (Pan, García-Díaz, and Valencia-García, 2024). This team participated in Tasks 1 and 2. To perform Task 1, their approach includes data preprocessing, sentiment feature, and fine-tuning of the pre-trained MarIA model. To undertake Task 2, they implement the same model as for Task 1, adding as a new feature the multiclass classification from Task 1.
- **UnibucAI** (Păduraru and Anghelina1, 2024). This team participated in Tasks 1, 2 and 3. A previous pre-processing of the texts has been carried out. For the prediction of Task 1, they used the RoBERTuito model and trained a layer of LSTMs. For Task 2 they used BETO as a model and used two approaches, one with a model trained by context and the other with a model for all contexts. For Task 3, they used the negative messages from Task 1 to create negative examples and two datasets to create positive examples.
- **NLP UNED MRES** (Sierra-Callau et al., 2024). This team participated in Tasks 1 and 2. In the training phase, they concatenated all messages from each user and split them into strings of less than 512 characters. The chosen model for the evaluation process was BETO UNCASED and no additional

preprocessing was going to be made (no stopwords removal or emoticon substitution).

- **ELiRF-VRAIN** (Casamayor et al., 2024). They explore three approaches. The first approach employs a classic machine learning algorithm, Support Vector Machines (SVM), which has shown satisfactory performance in long text classification tasks, serving as a benchmark for classical models. The second approach utilizes a pre-trained RoBERTa model, fine-tuned on the provided dataset and an expanded version created through data augmentation, leveraging Transformers for better adaptation to the task domain. The third approach, similar to the second, uses a pre-trained LongFormers model to capture more context due to its larger input size, and it is fine-tuned on the same datasets to enhance performance in long text classification.
- **VerbaNex AI** (Martinez et al., 2024). This team only participated in Task 1. Their approach involves data preprocessing applying data augmentation, lexical feature extraction, phonestheme embeddings, valence arousal dominance analysis (using NRC-Lexicon) and emotion detection. They used classical classifiers such as Decision Tree, Naïve Bayes or Multi-layer Perceptron to evaluate their performance.

6 Results

As mentioned in Section 2.4, tasks are evaluated according to how the task is defined. We evaluate a system according to its performance in terms of **absolute classification** and in terms of **early detection effectiveness** for classification tasks. Section 2.4 provides an overview of the evaluation metrics used for each task.

6.1 Task 1

Participant results and the baselines proposed are shown in Table 5 (absolute classification) and Table 6 (early detection).

In terms of absolute classification, the top-performing team, ELiRF-UPV, achieved the highest Macro F1 score of 0.874, surpassing the baseline models. Following the ranking, we have the teams UnibucAI and UNED-GELP, although these do not surpass the baseline RoBERTa Base. Notably, the models achieving the highest F1 results employ sentence-level relabeling of the data. For early detection (ERDE), the lowest value of ERDE30 (0.042) was achieved by the RoBERTa Base model, which surpasses the others. Among the participants, ELiRF-UPV had the best value, followed by UNED-GELP and UnibucAI. It is worth noting that UNED-GELP achieves the best ERDE5 value (0.138).

6.2 Task 2

This task involves two-level classification. The first is a multi-class classification setup where teams must detect if the user suffers from depression, anxiety or none of those disorders (same as task 1). The second is a multi-label classification, where teams had to determine the contexts from which the problem seemed to come. 6 teams have participated in this task, submitting 15 runs. Participant results and the baselines proposed are shown in Table 7 (absolute classification), Table 8 (early detection) and Table 9 (absolute classification for contexts).

Regarding the absolute classification, we observe the top-performing team, ELiRF-UPV, achieved the highest Macro F1 score of 0.874. This team outperformed the baseline models. The next highest-ranking teams, UnibucAI and Ixa-Med, also demonstrated strong performances but did not surpass the baseline RoBERTa Base. Remarkably, the models with the best f1 results apply sentence-level relabelling of the data. For early detection, the lowest value of ERDE30 (0.042) was obtained by the baseline model RoBERTa Base, which outperformed all other participant teams. ELiRF-UPV team obtained a high position in that ranking followed again by UnibucAI and Ixa-Med, however, teams like UC3M-DAD and UMUTeam (run 2) detected true positives with only one writing. However, the absolute classification for contexts completely rearranges the teams. UnibucAI Team dominated the top rankings with three submissions occupying the first, second, and fourth positions, respectively. UnibucAI (run 1) achieved the highest scores across most metrics, with a notable Macro-F1 of 0.268 and Micro-F1 of 0.291. UMUTeam (run 0) and ELiRF-UPV also get a macro-f1 higher than 0.2 outperforming baseline RoBERTa Base. Although ELiRF-

UPV gets the best values in multi-class classification, now its F1 scores were lower, indicating potential overfitting or imbalanced performance across classes. Top-performing teams like UnibucAI and UMUTeam had significant differences between their Macro and Micro scores, implying their models could handle the imbalance in the dataset better.

6.3 Task 3

This task has a binary classification setup where teams must detect if the user suffers or manifests symptoms of potential suicidal ideation. Labels will be 0 for "control" (negative, the user does not suffer from potential suicidal ideation) or 1 for "suffer" (positive). 4 teams have participated submitting 12 runs. Participant results and the baselines proposed are shown in Table 10 (absolute classification) and Table 11 (early detection). In terms of absolute classification, the top-performing team, UnibucAI, achieved the highest Macro F1 score of 0.534, surpassing the baseline models. The next team that surpassed baseline models is UNED-GELP with 0.456 of Macro F1. For early detection, the lowest value of ERDE5 (0.226) was obtained by the baseline model (all positives), which outperformed all other participant teams. Among the participants, the V team had the best value, followed by UNED-GELP and UnibucAI in ERDE5. The lowest value of ERDE30 (0.214) was obtained by the V team and also by the baseline model (all positives). UNED-GELP and UnibucAI followed Vteam in terms of ERDE30.

7 Discussion

Almost all participants applied large language models, being BETO and RoBERTuito were the most popular ones. The participating teams showed varying performance across tasks, with some models excelling in specific tasks. ELiRF-UPV team achieved the highest Macro F1 score of 0.874 in binary classification for both Task 1 and Task 2, surpassing baseline models. The baseline model RoBERTa Base achieved the lowest ERDE30 value (0.042) in early detection in both tasks, indicating strong performance in quickly detecting risks.

Task 2 also required a two-level classification: detecting the disorder (depression, anxiety, or none) and determining the context (e.g., addiction, emergency, family, work, social, other, none). In this case, the models struggled with context detection, highlighting the complexity and variability in the data.

Only four teams participated in Task 3, targeting suicidal ideation detection, indicating the challenge and sensitivity of the task. Performance metrics varied significantly, suggesting the need for more refined approaches and better training data.

Significant differences between Macro and Micro scores for top-performing teams indicate models' varying ability to handle class imbalance. Thus, there is potential overfitting or imbalanced performance across classes for some models, as indicated by lower F1 scores in multi-class classification despite high performance in binary classification tasks.

8 Conclusions

The MentalRiskES competition has made significant progress in detecting mental health disorders from social media data. Challenges remain, particularly in accurately identifying contexts and handling class imbalances.

Future iterations of the competition should focus on improving context detection and addressing class imbalance issues. Incorporating more diverse and representative training data could enhance model robustness and generalizability.

The ability to accurately detect mental health issues and suicidal ideation from social media can have profound implications for early intervention and support. Continued research and development in this area are crucial for advancing mental health detection technologies. In future editions, the targets could shift from disorder detection to symptom identification or suspicious behaviour annotation, as a way to provide more informative and explainable outputs and avoid black-box diagnostics.

Acknowledgments

Our sincere thanks to the organisers of the IberLEF workshop who, once again this year, have allowed us to organise the shared task.

This task has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21)

funded by Plan Nacional I+D+i from the Spanish Government.

References

Andreassen, C. S., S. Pallesen, and M. D. Griffiths. 2017. The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey. *Addictive behaviors*, 64:287–293.

Casamayor, A., V. Ahuir, A. Molina, and L.-F. Hurtado. 2024. ELiRF-VRAIN at MentalRiskES 2024: Using Longformers for Early Detection of Mental Disorders Risk. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Cedillo-Castelán, V. 2024. Suicidal ideation detection . In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Fernandez-Hernandez, J., H. Fabregat, A. Duque, L. Araujo, and J. Martinez-Romo. 2024. UNED-GELP at MentalRiskES 2024: Transformer-Based Encoders and Similarity Techniques for Early Risk Prediction of Mental Disorders. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

González, E. B. and M. T. Vidal. 2024. BUAP_01 at MentalRiskES 2024: Detection of depression and anxiety using features based on tagging. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

He, P., X. Liu, J. Gao, and W. Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Larrayoz, X., A. Casillas, M. Oronoz, and A. Pérez. 2024. Mental Disorder Detection in Spanish: hands on skewed class distribution to leverage training. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Losada, D. and F. Crestani. 2016. A test collection for research on depression and language use. volume 9822, pages 28–39, 09.

Mármol-Romero, A. M., A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Ráez. 2024. MentalRiskES: A new corpus for early detection of mental disorders in Spanish. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11204–11214, Torino, Italia, May. ELRA and ICCL.

Mármol-Romero, A. M., A. Moreno-Muñoz, F. M. Plaza-del Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, and A. Montejo-Raéz. 2023. Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish. *Procesamiento del Lenguaje Natural*, 71:329–350.

Martinez, E., J. Cuadrado, J. C. Martinez-Santos, and E. Puertas. 2024. Automated Detection of Depression and Anxiety Using Lexical and Phonestheme Features in Spanish Texts. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Muñoz-Muñoz, D., A. Marco-Perez, and D. Ramirez. 2024. Participation of UC3M-DAD on MentalRiskES task at IberLEF 2024. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Nakayama, H., T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Pan, R., J. A. García-Díaz, and R. Valencia-García. 2024. UMUTeam at MentalRiskES@IberLEF 2024: Using the Fine-Tuning Approach of Transformer-Based Models with Sentiment Feature for Early Detection of Mental Disorders. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2021. Overview of erisk

at clef 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, pages 864–887.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Păduraru, C. D. and I. M. Anghelina1. 2024. Early Risk Detection for Mental Health Disorders: UnibucAI at MentalRiskES 2024 . In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Sadeque, F., D. Xu, and S. Bethard. 2018. Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 495–503.

Shannon, H., K. Bush, P. J. Villeneuve, K. G. Hellemans, and S. Guimond. 2022. Problematic social media use in adolescents and young adults: Systematic review and meta-analysis. *JMIR Ment Health*, 9(4):e33450, Apr.

Sierra-Callau, M., M. Ángel Rodríguez-García, S. Montalvo-Herranzcand, and R. Martínez-Unanue. 2024. UNED_MRES team at MentalRiskES2024: Exploring hybrid approaches to detect Mental Disorder Risks in Social Media. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

A Participant Results

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	ELiRF-UPV	2	**0.890**	**0.875**	**0.880**	**0.874**
2	ELiRF-UPV	1	0.850	0.853	0.845	0.840
3	BaseLine - RoBERTa Base	2	0.853	0.840	0.843	0.834
4	ELiRF-UPV	0	0.848	0.840	0.838	0.833
5	UnibucAI	0	0.828	0.824	0.808	0.808
6	UnibucAI	1	0.820	0.802	0.798	0.795
7	UnibucAI	2	0.820	0.808	0.793	0.793
8	UNED-GELP	0	0.797	0.792	0.797	0.785
9	UNED-GELP	2	0.800	0.789	0.753	0.766
10	Ixa-Med	1	0.790	0.796	0.747	0.749
11	UNED-GELP	1	0.765	0.751	0.745	0.747
12	Ixa-Med	2	0.790	0.790	0.733	0.736
13	Ixa-Med	0	0.762	0.763	0.725	0.723
14	BaseLine - RoBERTa Large	1	0.670	0.786	0.708	0.682
15	UMUTeam	2	0.690	0.701	0.683	0.675
16	UMUTeam	0	0.630	0.728	0.662	0.640
17	BUAP_01	1	0.620	0.692	0.662	0.632
18	BaseLine - mDeBERTa	0	0.710	0.748	0.645	0.623
19	UC3M-DAD	0	0.578	0.727	0.647	0.601
20	UC3M-DAD	1	0.578	0.727	0.647	0.601
21	UC3M-DAD	2	0.578	0.727	0.647	0.601
22	NLP UNED MRES	0	0.557	0.644	0.620	0.561
23	BUAP_01	0	0.427	0.650	0.557	0.411
24	BUAP_01	2	0.393	0.348	0.352	0.348
25	VerbaNex AI	1	0.527	0.598	0.372	0.303
26	VerbaNex AI	2	0.527	0.598	0.372	0.303
27	VerbaNex AI	0	0.512	0.551	0.353	0.271
28	UMUTeam	1	0.515	0.712	0.355	0.269
29	NLP UNED MRES	1	0.352	0.564	0.402	0.264
30	NLP UNED MRES	2	0.318	0.664	0.383	0.237
31	Huerta	0	0.470	0.240	0.318	0.231
32	Huerta	1	0.470	0.240	0.318	0.231
33	Huerta	2	0.470	0.240	0.318	0.231

Table 5: Classification-based evaluation in Task 1. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	BaseLine - RoBERTa Base	2	0.162	**0.042**	3	0.969	**0.909**
2	ELiRF-UPV	2	0.405	0.045	8	0.891	0.845
3	ELiRF-UPV	0	0.453	0.060	9	0.875	0.801
4	UNED-GELP	0	**0.138**	0.065	2	0.984	0.880
5	UnibucAI	2	0.251	0.068	4	0.953	0.876
6	UnibucAI	1	0.279	0.069	4	0.953	0.874
7	ELiRF-UPV	1	0.414	0.074	7	0.906	0.816
8	UnibucAI	0	0.308	0.078	5	0.937	0.850
9	BaseLine - mDeBERTa	0	0.211	0.102	**1**	**1**	0.891
10	Ixa-Med	2	0.443	0.121	10	0.860	0.768
11	Ixa-Med	1	0.504	0.124	12	0.829	0.718
12	Ixa-Med	0	0.485	0.124	10	0.860	0.735
13	BaseLine - RoBERTa Large	1	0.205	0.133	**1**	**1**	0.811
14	BUAP_01	1	0.282	0.134	3	0.969	0.769
15	UNED-GELP	2	0.336	0.149	5	0.937	0.798
16	UNED-GELP	1	0.312	0.150	4	0.953	0.786
17	NLP UNED MRES	0	0.285	0.163	3	0.969	0.732
18	UC3M-DAD	0	0.227	0.165	**1**	**1**	0.756
19	UC3M-DAD	1	0.227	0.165	**1**	**1**	0.756
20	UC3M-DAD	2	0.227	0.165	**1**	**1**	0.756
21	UMUTeam	2	0.203	0.166	**1**	**1**	0.780
22	UMUTeam	0	0.593	0.194	11	0.844	0.629
23	NLP UNED MRES	1	0.341	0.209	2	0.984	0.695
24	NLP UNED MRES	2	0.427	0.225	4	0.953	0.657
25	BUAP_01	0	0.272	0.240	**1**	**1**	0.676
26	BUAP_01	2	0.363	0.359	**1**	**1**	0.522
27	VerbaNex AI	1	0.440	0.439	**1**	**1**	0.221
28	VerbaNex AI	2	0.440	0.439	**1**	**1**	0.221
29	VerbaNex AI	0	0.458	0.458	**1**	**1**	0.164
30	UMUTeam	1	0.501	0.501	80	0.154	0.013
31	Huerta	0	0.502	0.501	**1**	**1**	0.063
32	Huerta	1	0.502	0.501	**1**	**1**	0.063
33	Huerta	2	0.502	0.501	**1**	**1**	0.063

Table 6: Latency-based evaluation in Task 1. Metric ranking: ERDE30. In bold the best values for each metric are marked.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	ELiRF-UPV	0	**0.890**	**0.875**	**0.880**	**0.874**
2	BaseLine - RoBERTa Base	2	0.853	0.840	0.843	0.834
3	UnibucAI	2	0.828	0.824	0.808	0.808
4	UnibucAI	1	0.820	0.808	0.793	0.793
5	UnibucAI	0	0.820	0.808	0.793	0.793
6	Ixa-Med	1	0.790	0.796	0.747	0.749
7	Ixa-Med	2	0.790	0.790	0.733	0.736
8	Ixa-Med	0	0.762	0.763	0.725	0.723
9	BaseLine - RoBERTa Large	1	0.670	0.786	0.708	0.682
10	UMUTeam	2	0.690	0.701	0.683	0.675
11	UMUTeam	0	0.630	0.728	0.662	0.640
12	BaseLine - mDeBERTa	0	0.710	0.748	0.645	0.623
13	UC3M-DAD	0	0.578	0.727	0.647	0.601
14	UC3M-DAD	1	0.578	0.727	0.647	0.601
15	UC3M-DAD	2	0.578	0.727	0.647	0.601
16	NLP UNED MRES	0	0.550	0.630	0.608	0.550
17	NLP UNED MRES	1	0.550	0.630	0.608	0.550
18	NLP UNED MRES	2	0.550	0.630	0.608	0.550
19	UMUTeam	1	0.515	0.712	0.355	0.269

Table 7: Classification-based evaluation in Task 2. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	BaseLine - RoBERTa Base	2	**0.162**	**0.042**	3	0.969	**0.909**
2	ELiRF-UPV	0	0.405	0.045	8	0.891	0.845
3	UnibucAI	1	0.251	0.068	4	0.953	0.876
4	UnibucAI	0	0.251	0.068	4	0.953	0.876
5	UnibucAI	2	0.308	0.078	5	0.937	0.850
6	BaseLine - mDeBERTa	0	0.211	0.102	**1**	**1**	0.891
7	Ixa-Med	2	0.443	0.121	10	0.860	0.768
8	Ixa-Med	1	0.504	0.124	12	0.829	0.718
9	Ixa-Med	0	0.485	0.124	10	0.860	0.735
10	BaseLine - RoBERTa Large	1	0.205	0.133	**1**	**1**	0.811
11	NLP UNED MRES	0	0.253	0.156	2	0.984	0.750
12	NLP UNED MRES	1	0.253	0.156	2	0.984	0.750
13	NLP UNED MRES	2	0.253	0.156	2	0.984	0.750
14	UC3M-DAD	0	0.227	0.165	**1**	**1**	0.756
15	UC3M-DAD	1	0.227	0.165	**1**	**1**	0.756
16	UC3M-DAD	2	0.227	0.165	**1**	**1**	0.756
17	UMUTeam	2	0.203	0.166	**1**	**1**	0.780
18	UMUTeam	0	0.593	0.194	11	0.844	0.629
19	UMUTeam	1	0.501	0.501	80	0.154	0.013

Table 8: Latency-based evaluation in Task 2. Metric ranking: ERDE30. In bold the best values for each metric are marked.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**	Micro-P	Micro-R	Micro-F1
1	UnibucAI	1	0.022	0.194	**0.508**	**0.268**	**0.187**	**0.655**	**0.291**
2	UnibucAI	0	0.018	0.202	0.375	0.252	0.179	0.502	0.264
3	UMUTeam	0	0.007	0.166	0.408	0.224	0.165	0.647	0.263
4	UnibucAI	2	0.015	0.181	0.319	0.221	0.161	0.451	0.237
5	ELiRF-UPV	0	0.065	0.262	0.177	0.208	0.163	0.275	0.205
6	BaseLine - RoBERTa Base	2	0.075	**0.358**	0.168	0.181	0.183	0.314	0.232
7	UMUTeam	2	**0.077**	0.224	0.170	0.178	0.164	0.286	0.209
8	BaseLine - RoBERTa Large	1	0.070	0.356	0.139	0.164	0.166	0.275	0.207
9	UC3M-DAD	0	0.037	0.127	0.067	0.086	0.080	0.125	0.098
10	UC3M-DAD	1	0.037	0.127	0.067	0.086	0.080	0.125	0.098
11	UC3M-DAD	2	0.037	0.127	0.067	0.086	0.080	0.125	0.098
12	BaseLine - mDeBERTa	0	0.048	0.058	0.075	0.066	0.117	0.184	0.144
13	UMUTeam	1	0.000	0.169	0.026	0.044	0.023	0.039	0.029
14	Ixa-Med	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	Ixa-Med	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	Ixa-Med	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	NLP UNED MRES	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	NLP UNED MRES	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	NLP UNED MRES	2	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 9: Classification-based evaluation for multi-label in Task 2. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	UnibucAI	0	0.655	**0.556**	**0.539**	**0.534**
2	UnibucAI	1	0.600	0.499	0.499	0.496
3	UnibucAI	2	0.545	0.458	0.460	0.459
4	UNED-GELP	0	0.618	0.465	0.480	0.456
5	Baseline (all positives)	1	**0.691**	0.345	0.500	0.409
6	V team	0	**0.691**	0.345	0.500	0.409
7	V team	1	**0.691**	0.345	0.500	0.409
8	V team	2	**0.691**	0.345	0.500	0.409
9	UNED-GELP	1	0.673	0.343	0.487	0.402
10	UNED-GELP	2	0.382	0.454	0.455	0.382
11	Baseline (all negatives)	0	0.309	0.155	0.500	0.236
12	UNSL	1	0.309	0.155	0.500	0.236
13	UNSL	2	0.309	0.155	0.500	0.236
14	UNSL	0	0.291	0.148	0.471	0.225

Table 10: Classification-based evaluation for Task3. Metric ranking: Macro-F1. In bold the best values for each metric are marked.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	Baseline (all positives)	1	**0.226**	**0.214**	**1**	**1**	**0.817**
2	V team	0	0.261	**0.214**	**1**	**1**	**0.817**
3	V team	1	0.261	**0.214**	**1**	**1**	**0.817**
4	V team	2	0.261	**0.214**	**1**	**1**	**0.817**
5	UNED-GELP	0	0.326	0.215	**1**	**1**	**0.847**
6	UNED-GELP	1	0.344	0.232	2	**1**	0.796
7	UnibucAI	0	0.511	0.238	5	**1**	0.791
8	UnibucAI	1	0.654	0.317	10	**1**	0.729
9	UnibucAI	2	0.635	0.323	11	**1**	0.725
10	UNED-GELP	2	0.697	0.584	28	**1**	0.385
11	Baseline (all negatives)	0	0.691	0.691	nan	0	0
12	UNSL	1	0.691	0.691	nan	0	0
13	UNSL	2	0.691	0.691	nan	0	0
14	UNSL	0	0.703	0.703	nan	0	0

Table 11: Latency-based evaluation in Task 3. Metric ranking: ERDE30. In bold the best values for each metric are marked.