Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 329-350 recibido 09-06-2023 revisado 07-07-2023 aceptado 20-07-2023

Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish

Resumen de la tarea MentalRiskES en IberLEF 2023: 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]{amarmol, ammunoz, mdmolina, maite, laurena, amontejo}@ujaen.es,
[2]{flor.plaza@unibocconi.it}

Abstract: This paper presents the MentalRiskEs shared task organized at IberLEF 2023, as part of the 39th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023). The aim of this task is to promote the early detection of mental risk disorders in Spanish. We outline three detection tasks: Task 1 on eating disorders, Task 2 on depression, and Task 3 on an undisclosed disorder during the competition (anxiety) to observe the transfer of knowledge among the different disorders proposed. Furthermore, we asked participants to submit measurements of carbon emissions for their systems, emphasizing the need for sustainable NLP practices. In this first edition, 37 teams registered, 18 submitted results, and 16 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 2023, como parte de la 39ª 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 trastornos alimentarios, Tarea 2 para la depresión y Tarea 3 para identificar un trastorno que no desvelamos a los participantes (ansiedad) para observar la transferencia de conocimiento entre los distintos trastornos. Solicitamos medir emisiones de carbono para un desarrollo de modelos sostenible. En esta primera edición, 37 equipos se registraron, 18 enviaron predicciones y 16 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

According to a recent report by the World Health Organisation, there is 1 in every 8 people in the world suffering from a mental disorder (World Health Organization, 2022). The COVID-19 pandemic has raised the prevalence of anxiety and depression to more than 26% in just one year. Suicide is the fourth leading cause of death among 15-29 year-olds. The organisation considers that early identification is a crucial effective intervention to prevent these problems.

Consequently, there is a growing interest in detecting and identifying mental disorders in social media streams. This answers a demand from society due to the high increase in these problems among the population, in several kinds of mental risks: eating disorders, dysthymia, anxiety, depression, suicidal ideation, and others.

ISSN 1135-5948 DOI 10.26342/2023-71-26

In recent years, to analyse textual data and detect mental health problems such as depression, anxiety or suicidal ideation from user-generated content, researchers have increasingly turned to natural language processing (NLP) and deep learning. These computational methods offer promising opportunities for automated and scalable approaches to identifying people at risk or with mental health problems such as depression, anxiety or suicidal ideation from user-generated content. In fact, relevant evaluation campaigns like the Cross-Lingual Evaluation Forum (CLEF) have hosted during the last years the Early-Risk Identification task (eRisk) (Parapar et al., 2021). Unfortunately, these campaigns have focused mainly on English, leaving aside other languages, like Spanish.

MentalRiskEs (Mármol-Romero et al., 2023) is a novel task on early risk identification of mental disorders in Spanish comments from Telegram users organized within the Iberian Languages Evaluation Forum (IberLEF 2023) (Jiménez-Zafra, Rangel, and Montes-y Gómez, 2023). The task must be resolved as an online problem, that is, the participants must be able 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 participant approaches. For this first edition, the disorders considered are eating disorders (EDs), depression, and an unknown one which is intended to assess the robustness of approaches for new disorders not known a priori.

2 Tasks

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

2.1 Task 1. Eating disorders detection

1.a. Binary classification Detect if the user suffers from anorexia or bulimia. Labels are 0 for "control" (negative, the user does not suffer from ED) and 1 for "suffer" (positive).

- **suffer**: EDs are recognized by a persistent pattern of unhealthy eating or unhealthy dieting. It is an inappropriate eating behaviour and an obsession with weight control. 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.
- **control**: The user does not present evidence of suffering from the disorder.

1.b. Simple regression Provide a probability for the user to suffer anorexia or bulimia. A value of 0 means 100% negative and a value of 1 would be 100% positive.

2.2 Task 2. Depression detection

2.a. Binary classification Detect if the user suffers from depression. Labels are 0 for "control" (negative, the user does not suffer from depression) or 1 for "suffer" (positive).

- **suffer**: 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 (persistent sadness, low mood, and a lack of interest or pleasure in activities that were previously rewarding and pleasurable).
- **control** The user does not present evidence of suffering from the disorder.

2.b. Simple regression Provide a probability for the user to suffer depression. A value of 0 means 100% negative and a value of 1 would be 100% positive.

2.c. Multiclass classification Decide one among four different classes:

- **suffer+against**: A person who suffers from the disorder and seeks/offers help or information to get out of the disorder and overcome it. The person is against the disorder.
- **suffer+in favour**: A person who suffers from the disorder and encourages (seeks/provides information) other users to go deeper into the disorder. They are in favour of the disorder.
- **suffer+other**: A person suffering from the disorder and is not related to the above categories.

- **control**: A person is considered not to be suffering from the disorder when he/she does not show symptoms of suffering from it. They may be specialists in the subject who are dedicated to helping, people who have suffered from it in the past or people who bother other users or talk about a subject other than the disorder.

2.d. Multi-output regression For each of the previous classes, the system has to provide a probability of belonging to that class. These values, as in task 1.b., are interpreted as 0 for a 100% confidence of the system not assigning the user to a class, and 1 for a 100% probability of assigning the user to a class. Note that the sum of the four probabilities must be 1.

2.3 Task 3. Non-defined disorder detection

This is a binary classification (suffer, control) in which participants are encouraged to use the systems developed for subtasks 1.a, 1.b, 2.a, 2.b to identify a different disorder that is unknown to them but is related to the previous ones (ED and depression).

3.a. Binary classification Detect if a user suffers from an unknown disorder. Labels will be 0 for "control" or 1 for "suffer" (positive). Participants can use the systems developed for subtasks 1.a. and 2.a.

3.b. Simple regression Provide a probability for the user to suffer from the unknown disorder. A value of 0 means 100% negative and a value of 1 would be 100% positive.

For this task, participants can use the systems developed for subtasks 1.b. and 2.b.

2.4 Evaluation measures

Tasks are evaluated according to how the task is defined. We evaluate a system according to its performance in terms of **absolute classification** or in terms of **early detection effectiveness**. Besides, regression is evaluated on an error basis or on a ranking basis. Table 1 (Appendix A) summarizes the metrics computed for each task proposed.

2.4.1 Classification-based evaluation

This form of evaluation revolves around the binary decisions (binary or multi-class classification) taken for each user by the participating systems. This decision measures if a user has or does not have a risk of suffering a mental risk. To measure the tasks 1.a, 2.a, 2.c and 3.a we used classical metrics such as accuracy, macro-precision, macro-recall and macro-f1. This takes into account the final predictions of the system, once they know all the posts from each subject from the dataset. In order to rank the systems we chose the **macro-f1** metric.

2.4.2 Latency-based evaluation

We rely on the competition already established by eRisk (Parapar et al., 2021) to extract some metrics that measure the early detection of the positive subject from participating systems. To measure the tasks 1.a, 2.a, 2.c and 3.a we used early risk evaluation metrics such as ERDE (Losada and Crestani, 2016) (ERDE5 and ERDE30), latencyTP, speed and latency-weightedF1 (Sadeque, Xu, and Bethard, 2018). In order to rank the systems, due to the short length of the messages in our dataset, we consider that a larger number of messages are necessary to consider for early detection, so we apply the **ERDE30** metric.

About the early detection in multi-class classification (task 2.c.), we consider if a user is positive or not to measure systems.

2.4.3 Regression-based evaluation

This form of evaluation revolves around the score decisions (simple regression or multi-output regression) taken for each user by the participating systems. This score measures the level of risk that a user has. To measure the tasks 1.b, 2.b, 2.d and 3.b we used classical metrics such as RMSE and Pearson's coefficient. This takes into account the final predictions of the system, once they know all the posts from each subject from the dataset. In order to rank the systems, we consider the **RMSE** metric. We consider the mean of the measures calculated for each label in multi-output regression (task 2.d).

2.4.4 Ranking-based and multi-output regression evaluation

To measure the performance of the system in determining the level of severity of certain users being at risk of suffering from a mental disorder in comparison with others, we apply a ranking-based evaluation. In tasks 1.b, 2.b, 2.d, and 3.b we used the Precision@K, which is the Precision at top-k (users with the highest scores). This measures how many risk

subjects are present in the top-k recommendations of your system. The possible values of k are 5, 10, 20, or 30. In order to rank the systems, we consider the **p@30** metric at round 25. We consider the mean of the measures calculated for each label in multi-output regression (task 2.d).

2.4.5 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. In Appendix A.1 this section is described in more detail.

3 Dataset

We have used the messaging platform Telegram[1] to collect messages from users suffering from mental disorders. Prolific[2] to search for annotators and Doccano (Nakayama et al., 2018) to perform the annotation process.

3.1 Compilation

We used data from some public groups on the Telegram messaging application. Telegram via the application allows downloading messages from public groups. This data was downloaded in May 2022. Table 2, in Appendix B, shows the names of the public groups used.

3.2 Curation

URLs, hashtags, and bold-style text are replaced with targets, while messages containing less than three tokens are excluded. Moreover, emojis are converted into their corresponding text representations. Additionally, to anonymize the messages, names, aliases, and telephone numbers are removed.

Then, we removed subjects whose number of messages fell below or exceeded specified limits. For ED, the minimum limit was set at 10 messages, while the maximum limit was 50. In other cases, the maximum limit was 100. If a subject exceeded these limits, their messages were truncated to the most recent 50 or 100 messages accordingly. Additionally, we carefully selected a specific number of users to ensure equal representation.

[1]https://telegram.org/

[2]https://www.prolific.co/

3.3 Annotation

We used Prolific and Doccano for annotating the collected data. Prolific helped us to recruit annotators and Doccano is an open-source text annotation tool that allows annotators to do their work.

An annotation guide was developed for each of the datasets. The annotation guides provided annotators with examples of each label, a list of frequently asked questions, and a graphical outline to facilitate understanding. Furthermore, a user manual was developed to guide the use of Doccano. Once the annotation guidelines were set up and the software was configured, the annotation took approximately four months to complete.

3.4 Agreement

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 the 10 annotators. The final results are in Table 3, Appendix C.

After annotating the corpora, we decided that it was more coherent to link the respective classes to the risk of a user suffering from an ED as there were hardly any subjects for the "Suffer+other" and "Suffer+against".

3.5 Dataset statistics

A total of three datasets are presented, encompassing ED, depression, and anxiety. The first and the last contain subjects who can be considered at risk for a disorder and those who are not, while the depressive dataset contains control subjects and subjects suffering from the disorder divided into three categories. 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 the systems, and (3) test: to test the systems. In total, there are 335, 334, and 150 subjects for ED, depression and anxiety, respectively. The distribution of subjects in the sets and tasks can be seen in Table 4, Appendix C.

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 (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 were sent. The attributes for each JSON were: (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 MentalRiskEs corpus, we performed experiments using three different Transformer-based models. We experimented 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,[3] RoBERTa Large[4] and mDeBERTa[5] are available at the HuggingFace models' hub.[6]

Details about different configurations of the models and the training process are shown in Table 5, Appendix D.1. For all experiments, we trained using the training set, used the trial set for early stopping, and evaluated using the test set.

4.1 Binary classification

In the HuggingFace transformer training arguments, the number of labels was set to 2 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. The results and the epochs, in which each model was trained, are depicted in Table 6, Appendix D.2.

4.2 Simple regression

In the HuggingFace transformer training arguments, the number of labels was set to 1 and the problem type was set to regression. Early stopping was set to stop when the lowest value in the RMSE metric was reached. The results and the number of epochs, in which each model was trained, are depicted in Table 7, Appendix D.2.

4.3 Multi-class classification

In the HuggingFace transformer training arguments, the number of labels was set to 4 and the problem type was set to single-label classification. Early stopping was set to stop when the highest value in the macro-averaged F1 score was reached. The results and the epochs, in which each model was trained, are depicted in Table 8, Appendix D.2.

4.4 Multi-output regression

In the HuggingFace transformer training arguments, the number of labels was set to 4 and the problem type was set to multi-label classification. Early stopping was set to stop when the lowest value in the mean of the RMSE metric calculated for each class was reached. The results and the number of epochs, in which each model was trained, are depicted in Table 9, Appendix D.2.

5 Participant approaches

A total of 37 teams from 8 countries (Spain, Ireland, Mexico, Chile, Canada, Colombia, China, and Argentina) signed up for MentalRiskES 2023. Among them, 18 teams submitted runs for Task 1, 28 for Task 2, and 8 for Task 3. Each team had the chance to submit a maximum of 3 runs, demonstrating their expertise and strategies in the challenge. In the following, we describe the approaches of the team that participated in the competition:

- **CIMAT-NLP** (García Santiago, Sánchez-Vega, and López-Monroy, 2023). This team participated in all tasks using RoBERTuito, a Spanish transformer model trained on textual data. They pursued two distinct approaches. The first approach involved aggregating messages of a fixed size into packages. In the second approach, they introduced data augmentation techniques during training. For task 3, they relied on the system developed for tasks 1 and 2, combining an ensemble of both models.
- **CIMAT-NLP-GTO** (Echeverría-Barú, Sanchez-Vega, and Pastor

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

López-Monroy, 2023). This team participated in all tasks. They explore both, a classical approach and a transformer-based. The former is based in a TF-IDF representation of the user's history over n-grams of characters; the resulting vectors are passed to a Naïve Bayes algorithm. The second approach trains, with different seeds, pretrained versions of RoBERTuito and the encoding of the user's history is varied, taking different hidden states from the neural network and combining them before a final feed-forward network.

- **GetitDone** (Hu and Zhou, 2023). They participated in task 2.a. and uses BERT, pre-trained on a Spanish corpus. For the tokenizer part, they use the default setting in the pre-trained model RoBERTuito-sentiment-analysis (with several heads, attention, ffn, etc.). In the classification part, they use a three-class sentiment analysis classifier. Thus, the sum of neutral and positive probabilities is considered the non-depressed probability, and the negative probability is considered the depressed probability.

- **I2C-UHU** (Vázquez Ramos et al., 2023). They participated in tasks 1.a., 1.b. and 2.c. This team tested several pre-trained models, like BERT, DeBERTa or RoBERTa-BNE. A strategy for finding the best hyperparameters is applied, along with data augmentation by two-step back-translation (Spanish to English, English to French and French to Spanish). The training is performed at the message level.

- **NLP-UNED** (Fabregat et al., 2023). This team participated in all tasks. The algorithm applied is Approximate Nearest Neighbours (ANN) over representations of each message with a Universal Sentence Encoder (USE). Messages are, prior to classification, relabeled in order to improve the separability of resulting clusters.

- **NLPUTB** (Martinez et al., 2023). This team participated in subtasks 2.a. and 2.b. To accomplish this task, their approach involves data pre-processing, lexical feature extraction, and phonesthemes embedding which encodes phonetic information using the RoBERTuito model to capture contextual representations. As classifiers, they leveraged traditional machine learning classifiers such as Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, and k-Nearest Neighbors.

- **PLN-CMM** (Guerra et al., 2023). They participated in tasks 1.a., 1.b., 2.b., and 2.d. They use a classical approach using lemmatization, stop word removal, and bigrams or trigrams weighting with TF-IDF. For task 1.a., the used classifiers are Linear SVC, RBF SVC, KNN, XGB, and MultinomialNB. For tasks 1.b., 2.b. and 2d the used regressors were SGD, Ridge, Linear, Gradient Boost, and Random Forest. But, only in task 2.d., do they use trigrams with TF-IDF.

- **SINAI-SELA** (González-Silot, Martínez-Cámara, and Ureña-López, 2023). They participated in subtask 2.a., after a previous analysis they concluded that all the messages are needed to identify if there is a mental disorder and used two transformed-based language models, in one they used fine-tuning with LM BETO Emotion Analysis, and in the other one, they explored the possibility of adding emotion knowledge to the model.

- **SPIN** (Zubiaga and Justo, 2023). They participated in subtasks 2.a. and 2.b. For the binary one, all messages are used to represent the user. A reference text, like "I'm depressed", is also encoded. Cosine distances from the user's messages and the reference one are used to determine the binary decision. For the multiclass one, when the former binary approach predicts a positive, a similar approach is done, but over six different reference expressions. Cosine similarities are computed and combined with statistical information over them to generate a final vector that is fed to a feed-forward network for final classification.

- **TextualTherapist** (Fernández-Hernández et al., 2023). This team participated in subtask 2.a. The extract from the user's history has a lot of

different features (PoS, readability and toxicity metrics, emotions, mapping to psychological survey indicators, deep encodings, and LIWC vectors, among others). All those features are combined into a final vector that passes to classical machine learning algorithms, like Random Forest, Logistic Regression or Light Gradient Boosting Machine. A feature selection process over model performance is applied.

- **UMUTeam** (Pan, García-Díaz, and Valencia-García, 2023). This team participated in all subtasks of Task 1 and Subtask 2.a and 2.b. For the subtasks involving binary classification, they proposed a fine-tuning approach using both pre-trained monolingual and multilingual models. Additionally, they experimented with an ensemble learning technique. In tackling the regression problems, they leveraged the classification models and incorporated a softmax transformation to the output to obtain the probability of experiencing distress. Throughout their work, they explored various Transformer models such as BETO, ALBETO, DistilBETO, MarIA and XLM-RoBERTa.
- **DepNLP UC3M GURUDASI** (Sánchez-Viloria et al., 2023). This team participated in all subtasks of Task 2, using a combination of traditional machine learning and deep learning techniques. They processed and augmented the dataset by grouping messages per user and combining them into a single string. To increase training data and simulate early detection, they included observations with only half of the messages. They explored two approaches: fine-tuning a RoBERTa model (pre-trained on Spanish texts) and training a standard machine learning regressor using sentence embeddings from user messages as features.
- **UNSL** (Thompson and Errecalde, 2023). This team participated in tasks 1.a. and 2.a. Their approach involved leveraging the Transformer model BETO and using a decision policy guided by an early detection framework's predefined criteria. One of their models introduced an expanded vocabulary comprising crucial words specific to each task. Furthermore, they incorporated a decision policy that considered the model's prediction history during user evaluation.
- **UPM** (Rujas et al., 2023). This team participated in tasks 1.a. and 1.b. They carry on a large preprocessing of the data and topic modeling (BERTopic). They use a BERT-based model (BETO) that is fine-tuned and is passed different types of input which integrate temporal data (time or date of the message), preprocessed text messages, and the topic.
- **VICOM-nlp** (Turón et al., 2023). This team participated in tasks 1.a. and 2.a. They relabeled the data at the message level. A new dataset of sub-streams is created and labelled using a "confidence" value and then, a BERT-based model is fine-tuned.
- **Xabi IXA** (Larrayoz et al., 2023). This team proposed an approach for text classification using an encoder to transform text messages into numerical vectors, which are then fed into a feed-forward neural network. Two encoders, Dynamic Aggregation of Networks (DAN) and Sentence-BERT (SBERT) are used to generate input for the FFNN. The training process involves assigning user-level labels to posts and applying a weighted cross-entropy loss function to handle imbalanced data and prioritize false positives or false negatives.

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. Besides, regression tasks are evaluated on an error basis or on a ranking basis. Section 2.4 provides an overview of the evaluation metrics used for each task.

6.1 Task 1

Task 1.a. This task involves a binary classification setup where teams must detect if the user suffers from an ED. 10 teams have participated in this subtask and there are submitted 22 runs. Participant results

and the baselines proposed are shown in Table 10 (absolute classification) and Table 11 (early detection). About the first one, 6 teams have surpassed the baseline DeBERTa and RoBERTa Large. One of them, CIMAT-NLP-GTO obtained the highest value in all the metrics but the difference with the top 3 teams (UMUTeam and UNSL) is very low. Regarding the early detection, we can observe that CIMAT-NLP-GTO is again the team with the lowest value of ERDE30. UNSL, VICOM-nlp and CIMAT-NLP surpassed the DeBERTa baseline. Similarities among the teams include the use of transformer-based models, with CIMAT-NLP-GTO, UNSL, UMUTeam and VICOM-NLP using models like BETO or MarIA. Some teams also used ensemble approaches to improve performance. However, there were differences in their architectures and dataset preprocessing methods. CIMAT-NLP-GTO used a combination of Bag-of-Characters models and Transformers, UNSL extended the vocabulary of BETO with an augmented dataset, UMUTeam used MarIA and pysentimiento (Pérez, Giudici, and Luque, 2021) with ensembles, and VICOM-NLP applied data relabeling at the post level.

Task 1.b. Task 1.b. consists in the determination of the probability of suffering from ED, so it was evaluated as a simple regression problem using the RMSE metric as a reference measure. Eight teams submitted a total of 17 different runs. The results of the evaluation for this task are shown in Table 12. None of the submitted predictions exhibited better performance than that of the RoBERTa Base baseline (0.178) being the closest one the run 1 by the CIMAT-NLP-GTO team (0.192). If we look at Pearson's coefficient, this baseline achieves an impressive value of 0.906, which makes the model's predictions almost match those of humans' estimations. The third best value was reported by another baseline model: RoBERTa Large. CIMAT-NLP-GTO also holds the fourth-best position with its run 2. These runs are based on an ensemble of 5 transformer-based models. The second (run 2) applies also a data augmentation process. Close values are reported by runs submitted by teams CIMAT-NLP. Ranking-based evaluation is shown in Table 13. According to the reference metric, P@30 in round 25, the baseline model RoBERTa Large was the best with a value of 0.900, followed by CIMAT-NLP-GTO's run 2 (0.867), with the baseline BeBERTa and run 0 of CIMAT-NLP-GTO reporting same P@30 values. This last system is based on character n-grams and classical TF-IDF weighting to feed a Naïve Bayes system. This makes us think that a special vocabulary is used by people suffering from this disorder.

6.2 Task 2

Task 2.a. This task involves a binary classification setup where teams must detect if the users suffer from depression. 14 teams have participated in this subtask, submitting 30 runs. Participant results and the baselines proposed are shown in Table 14 (absolute classification) and Table 15 (early detection). Regarding the absolute classification, the top 6 teams have surpassed the baseline RoBERTa Large model. In particular, UMUTeam, UNSL, and TexualTherapist achieved the highest Macro-F1 scores in the task, with a small margin difference between them (0.737-0.729). The early detection evaluation appears to be more challenging, as only SINAI-SELA and UNSL managed to surpass the DeBERTa and RoBERTa Large baselines in terms of ERDE30 (the official ranking metric), achieving values of 0.140 and 0.148 respectively. The RoBERTa Base baseline ranks in the 10^{th} position, with 23 teams unable to surpass it. UMUTeam and SINAI-SELA delve into the integration of emotional features. TextualTherapists also explores the inclusion of emotion features alongside PoS, toxicity metrics, and more. The incorporation of features, particularly emotions, seems to enhance the detection capability of the models on this task. UNSL, using the monolingual BETO, achieves the best results and uses a decision policy that takes into account the model's prediction history during user evaluation.

Task 2.b. This task approaches depression detection as a simple regression problem. Systems must output a probability for the individual to suffer from depression or not. Seven participating teams submitted a total of 16 different runs. The evaluation metrics on simple regression for these runs along with those of the proposed baselines are reported in Table 16. The runs are ordered according to their RMSE error. It can be seen that as for task 1.b., none of the submitted predictions was able to overcome the RoBERTa

Base baseline (0.277) being the closest one the run 1 by the CIMAT-NLP-GTO team (0.292). If we look at Pearson's coefficient, this baseline achieves a remarkable value of 0.770, which makes the model's predictions very close to humans' estimations. CIMAT-NLP-GTO also holds the third-best position. These runs are based on an ensemble of 5 transformer-based models. The second (run 2) applies also a data augmentation process. Compared to a similar task, 1.b., it is clear that detecting depression is more difficult than detecting ED. Ranking-based evaluation is shown in Table 17. When looking at the reference metric, P@30 in round 25, the best system was that of the run 0 by the PLN-CMM team, with a superior value of 0.600. As can be noticed, precisions fall quickly to very low values, despite a comparable RMSE value being obtained. This tells us that only a few systems are able to, within an RMSE error close to 0.3, produce scores (probabilities) that can be trusted as the probability of suffering from depression or not.

Task 2.c. This task focuses on a multi-class classification scenario, where participants have to detect one among the following four classes (suffer+against, suffer+in favour, suffer+other, and control). A total of 5 teams have participated in this subtask, submitting 10 runs. The performance of participants and the proposed baselines are shown in Table 18 (absolute classification) and Table 19 (early detection). In terms of absolute classification, this task proves to be quite challenging, as none of the participants outperformed the baseline RoBERTa large, which achieved a Macro-F1 score of 0.360. In fact, the obtained results are notably low when compared to other tasks. NLP-UNED achieved the second position with a Macro-F1 score of 0.358 in run 1 and 0.339 in run 0. This team applied the ANN algorithm over representations of each message with USE and relabeled the messages in order to improve the separability of resulting clusters. The baseline DeBERTa ranked third with a Macro-F1 score of 0.293. PLN-CNN occupied the fifth position, while four teams fell short of surpassing the baseline RoBERTa Base, which achieved the sixth position. The early detection evaluation based on the ERDE30 has a common pattern. None of the participants exceeds the performance of the baseline DeBERTa. I2C-UHU achieved the second position in the ranking with an ERDE30 score of 0.198, while NLP-UNED takes the 3rd position with a score of 0.203. The best team, I2C-UHU, applied data augmentation with back-translation and fine-tuned a RoBERTa base model trained on Spanish texts.

Task 2.d. This task involves a multi-output regression setup where teams have to detect a score for each class available. A total of 2 teams have participated in this subtask, submitting a total of 4 runs. Participant results and the baselines proposed are shown in Table 20 (multi-output regression) and Table 21 (ranking-based). Regarding the regression, we can observe that neither DepNLP UC3M GURUDASI nor NLP-CMM teams have surpassed DeBERTa Baseline. However, in the ranking table, the first team obtain the highest values surpassing all baselines. DepNLP UC3M GURUDASI used a fine-tuning RoBERTa model and NLP-CMM used trigrams with TF-IDF and classical machine learning algorithms.

6.3 Task 3

Task 3.a. This task involves a binary classification setup where teams must detect if the users suffer from an unknown disorder. A total of 4 teams have participated, submitting 10 runs. Participant results and the baselines proposed are shown in Table 22 (Binary classification) and Table 23 (Latency evaluation). Only the CIMAT-NLP-GTO team has surpassed the baseline DeBERTa model achieving the highest Macro-F1 scores in the task (0.740). NLP-UNED achieved the 3rd position with a Macro-F1 score of 0.650 in run 1, ahead of the RoBERTa Large baseline. Between this Baseline and the 8th position of RoBERTa Base Baseline, three teams (CIMAT-NLP, NLP-UNED and CIMAT-NLP-GTP) achieved 0.614, 0.595 and 0.593 Macro-F1 score, respectively. The early detection evaluation appears to be more challenging, as none team managed to surpass the DeBERTa and RoBERTa Large baselines in terms of ERDE30 (the official ranking metric), and only the CIMAT-NLP-GTO team outperformed the Baseline RoBERTa Base by achieving a value of 0.188. The RoBERTa Base baseline ranks in the 4th position, with 3 teams unable to surpass it. The best team, CIMAT-NLP-GTO, implements an ensemble of 10 transformer-based models. The first five models applied data

augmentation and the second set of five models applied only the training set of Task 2.

Task 3.b. This task provides a probability for the user to suffer from the unknown disorder. A total of 4 teams have participated in this subtask, submitting 10 runs. Participant results and the baselines proposed are shown in Table 24 (Simple regression) and Table 25 (Ranking-based). In terms of simple regression, this task proves to be quite challenging, as none of the participants outperformed the baselines RoBERTa Base and DeBERTa. Between these Baselines, CIMAT-NLP-GTO and CIMAT-NLP have achieved 0.329 and 0.332 in terms of RMSE. In terms of simple regression based on P@30, the same as in the previous task, none of the participants outperformed the same baselines. The 3rd position is for CIMAT-NLP-GTO team achieving a 0.667. The next team is UPM with 0.633 and the last rank is BaseLine RoBERTa Large. For the ranking-based evaluation, the best team, CIMAT-NLP-GTO, implements the same model explained in task 3.a. In the simple regression evaluation, the same team in run 1, used an ensemble of 5 transformer-based models being each model trained with the training set of Task 2.

7 Discussion

Most of the approaches considered sampling at the subject level (thus, concatenating messages in the user's history). Few of them were sampled at message level (like VICOMnlp or UMUTeam), with performing results. Ten of the participants opted for fine-tuning pre-trained models like BETO or RoBERTuito. This last one seems to be a very good choice when dealing with mental disorder detection. Such approaches were among the top-ranked ones. Two teams (NLP-UNED and Xabi IXA) explored the use of sentence encoders, but the results show that this kind of Transformers has room for improvement. The results obtained by TextualTherapists demonstrate that intensive feature-engineered methods come closer to end-to-end solutions, at least in depression detection. In this sense, four teams applied classical machine learning algorithms in their approaches (like Random Forest, Naïve Bayes, or Support Vector Machines, among others).

It can be drawn from the results that there is no one-fits-all solution. The different approaches and attempts overcome by participating teams perform differently depending on the target disorder. Eating disorders seem to hold their own terminology, so word-based and character-based vectors resulted in very performing systems. Compared to depression, that disorder was found easier to predict by different approaches.

All submissions were accompanied by efficiency measurements. Although the analysis of those metrics has not been included in this report, we must highlight that the system showing the best balance between efficiency and performance was by UMUTeam, with a very low carbon footprint and competitive performance results.

8 Conclusions

This new task at the IberLEF forum has had a significant response from the scientific community, with 16 teams participating from all around the world, despite the complexity of the submission system and the participation requirements. The variety of the disorders considered and the profusion of evaluation approaches have leveraged the knowledge of automatic detection of mental disorders in social networks in the Spanish language. Although deep-learning models are the preferred ones, there is still room for alternative and classical solutions with competitive performances. Preprocessing the data are among the most challenging tasks, as early detection in a stream of messages poses new and creative solutions to define what is a sample (a single post? a window of messages? the full history of the user?) and how it is labelled. We plan to organize future editions of this lab, as automatic detection of mental disorders seems a promising application of natural language technologies and has a significant impact on society.

Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

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

Echeverría-Barú, F., F. Sanchez-Vega, and A. Pastor López-Monroy. 2023. Early Detection of Mental Disorders in Spanish Telegram Messages using Bag of Characters and BERT Models. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Fabregat, H., A. Duque, L. Araujo, and J. Martinez-Romo. 2023. NLP-UNED at MentalRiskES 2023: Approximate Nearest Neighbors for Identifying Health Disorders. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

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.

Fernández-Hernández, A., R. Moreno-Sánchez, J. Viosca-Ros, R. Enrique-Guillén, N. P. Cruz-Díaz, and S. M. Jiménez-Zafra. 2023. TextualTherapists at MentalRiskES-IberLEF2023: Early Detection of Depression using a User-level Feature-based Machine Learning Approach. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

García Santiago, M.-d.-J., F. Sánchez-Vega, and A. P. López-Monroy. 2023. Improving Transformer by Instance Packaging for Mental Illnesses identification. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

González-Silot, S., E. Martínez-Cámara, and L. A. Ureña-López. 2023. SINAI at MentalRisk: Using Emotions for Detecting Depression. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Guerra, R., B. Pizarro, C. Aracena, C. Muñoz-Castro, A. Carvallo, M. Rojas, and J. Dunstan. 2023. CMM PLN at MentalRiskES: A Traditional Machine Learning Approach for Detection of Eating Disorders and Depression in Chat Messages. 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*.

Hu, C. and X. Zhou. 2023. Mental Disorders Detection with Immediate Message using RoBERTa. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Jiménez-Zafra, S. M., F. Rangel, and M. Montes-y Gómez. 2023. Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org*.

Larrayoz, X., N. Lebeña, A. Casillas, and P. Alicia. 2023. Eating disorders detection by means of deep learning. 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.

Martinez, E., J. Cuadrado, D. Peña, J. C. Martinez-Santos, and E. Puertas. 2023. Automated Depression Detection in Text Data: Leveraging Lexical Features, phonesthemes Embedding, and RoBERTa Transformer Model. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

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. 2023. Overview of MentalriskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish. *Procesamiento del Lenguaje Natural*, 71.

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. 2023. UMUTeam at MentalRiskES2023@IberLEF: Transformer and

Ensemble Learning Models for Early Detection of Eating Disorders and Depression. 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.

Pérez, J. M., J. C. Giudici, and F. Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Rujas, M., B. Merino-Barbancho, P. Arroyo, and G. Fico. 2023. Development of a Natural Language Processing-Based System for Characterizing Eating Disorders. 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.

Sánchez-Viloria, S., D. Peix-del Río, R. Bermúdez-Cabo, G. A. Arrojo-Fuentes, and I. Segura-Bedmar. 2023. A Framework for Identifying Depression on Social Media: MentalRiskES@IberLEF 2023. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Thompson, H. and M. Errecalde. 2023. Early Detection of Depression and Eating Disorders in Spanish: UNSL at MentalRiskES 2023. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Turón, P., D. Cabestany, N. Pérez, and M. Cuadros. 2023. Text Classification For Early Detection of Eating Disorders and Depression in Spanish. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Vázquez Ramos, L., C. Moreno García, J. Mata Vázquez, and V. Pachón Álvarez. 2023. I2C-UHU at MentalRiskES 2023: Detecting and Identifying Mental Disorder Risks in Social Media using Transformer-Based Models. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

World Health Organization. 2022. Mental disorders. https://www.who.int/news-room/fact-sheets/detail/mental-disorders, June. Accessed: 2023-02-10.

Zubiaga, I. and R. Justo. 2023. SPIN at MentalRiskES 2023: Transformer-Based Model for Real-Life Depression Detection in Messaging Apps. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

A Evaluation metrics

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. The reference metric (for submission ranking) for that evaluation is in bold.

A.1 Efficiency metrics

We want to recognize those systems that are able to perform the task with minimal demand for resources. This will allow us to, for instance, identify those technologies that could run on a mobile device or a personal computer, along with those with the lowest carbon footprint. To this end, each final prediction calculated contain the following information: (a) minimum, maximum, mean and variance of time to make a prediction, (b) minimum, maximum, mean and variance of CO2 emissions generated when making a prediction, (c) minimum, maximum, mean and variance of energy per CPU or/and GPU (kW) used when making a prediction, (d) minimum, maximum, mean and variance of energy used per RAM (kW) when making a prediction, (e) minimum, maximum, mean and variance of sum of CPU energy, GPU energy and RAM energy (kW) consumed, (f) number of CPU or/and GPU used and their models and the total ram size needed.

Participants used the CodeCarbon package[7] which enables them to track emissions, measured as kilograms of CO2-equivalents (CO2eq) in order to estimate the carbon footprint of their systems predictions.

B Telegram groups

Table 2 show the titles or names and *groupsnames* or usernames from a group which we used to extract the messages to create the dataset. It is important to consider that

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

Subtasks	Evaluation perspective	Metrics
1.a., 2.a., 3.a.	Absolute binary classification	Accuracy, Macro-P, Macro-R **Macro-F1**
1.a., 2.a., 3.a.	Early detection in binary classification	ERDE5, **ERDE30**, latencyTP, speed, latency-weightedF1
1.b., 2.b., 3.b.	Simple regression	**RMSE**, Pearson's coefficient
1.b., 2.b., 3.b.	Ranking on simple regression	P@5, P@10, P@20, **P@30**
2.c.	Absolute multi-class classification	Accuracy, Macro-P, Macro-R **Macro-F1**
2.c.	Early detection in multi-class classification	ERDE5, **ERDE30**, latencyTP, speed, latency-weightedF1
2.d.	Multi-output regression	RMSE, Pearson's coefficient (both for each class), **RMSE mean**, Pearson's coefficient mean
2.d.	Ranking on multi-output regression	P@5, P@10, P@20, **P@30**

Table 1: Metrics used in the evaluation of submissions to MentalRiskEs subtasks.

messages' dates could be very different in the ED dataset for different subjects due to we needed more than one group to create it. Some messages or groups may have been deleted.

C Corpus

This section describes the number of subjects and messages existing in each set (trial, train, and test) and the tasks proposed. Table 4 show a summary of the subjects' distribution and messages' distribution in each set and task. Moreover, Table 3 shows the Cohen's kappa scores for each dataset.

Dataset	4 labels	2 labels
ED	0.185	0.249
Depression	0.316	0.521
Anxiety	-	0.449

Table 3: Cohen's kappa scores for each dataset and with a binary classification or multi-class classification.

D Baselines

This appendix presents the parameters established in the baseline experiments as well as the most relevant results in the tasks proposed.

D.1 Baseline hyper-parameters

The experiments with Transformer used default hyper-parameters, however, we apply a fine-tuning that is specified in Table 5 and added a TrainerCallback to handle early stopping. All the training and evaluation experiments were performed on a node equipped with 2 NVIDIA V100 servers. In these GPUs, each Volta V100 has a memory of 32GB, and the number of cores it provides is 5,120 CuDA FP32 cores and 640 Tensor cores.

The epoch from each experiment is established in the next subsection because it was determined by the early stopping callback in the training phase.

Hyperparameters	Value
Learning Rate	5e-5
Weight Decay	0
Batch size	8
Seed	42
Max length	512

Table 5: Baselines training details.

D.2 Baselines experiments

In Table 6 are the epochs used in each task about binary classification (Task 1.a., Task 2.a. and Task 3.a.) and each model next to the final macro-f1 score (rank metric) obtained in the test phase. Table 7 is the same information but about simple regression tasks (Task 1.b., Task 2.b. and Task 3.b.) with the RMSE metric. In Table 8 and

Mental disorder	Group name	Telegram group
ED	The voice filtro	anaymiarex
	Anorexia y bulimia	e12345gk
	Anorexic boy	anorexicovivir
	Musculación Ibérica	gimnasio
	Grupo de Apoyo para Bajar de Peso	grupodeapoyoparabajardepeso
	Comida Sana	_comida_sana
	Chat free Comer Sano y Saludable	comersanok
	Bajar de peso sanamente	baja_de_peso_sanamente
Depression	Superando la depresión	incomprendidos
Anxiety	Aprendiendo a vivir con la ansiedad	enluchaconstante

Table 2: Telegram groups used to create the corpus.

		ED		Depression		Anxiety	
		Subjs.	Msgs.	Subjs.	Msgs.	Subjs.	Msgs.
Trial	suffer+favour	5	161	2	35	-	
	suffer+against			2	136		
	suffer+other			2	126		
	control	5	228	4	327		
	Total	10	389	10	624		
Train	suffer+favour	74	2,532	44	1,524	-	
	suffer+against			44	1,457		
	suffer+other			6	132		
	control	101	3,399	81	3,135		
	Total	175	5,931	175	6,248		
Test	suffer+favour	64	1,220	32	1,154	93	3,298
	suffer+against			31	1,042		
	suffer+other			5	143		
	control	86	2,959	81	2825	57	2,222
	Total	150	4,179	149	5,164	150	5,520

Table 4: Number of subjects and messages' distribution by label and by set.

Table 9 are the epochs used in multi-class classification tasks and multi-output regression tasks, task 2.c. and task 2.d. respectively. The first is rank by macro-f1 metric and the last one is rank by the mean of the RMSE values calculated before for each class ("suffer+in favour", "suffer+against", "suffer+other" and "control").

Task	Model	Epoch	Macro-F1
Task 1	DeBERTa	2	0.813
	RoBERTa Large	4	0.813
	RoBERTa Base	3	0.694
Task 2	DeBERTa	2	0.642
	RoBERTa Large	4	0.690
	RoBERTa Base	3	0.605
Task 3	DeBERTa	5	0.693
	RoBERTa Large	3	0.630
	RoBERTa Base	3	0.553

Table 6: Results for binary classification.

Task	Model	Epoch	RMSE
Task 1	DeBERTa	6	0.231
	RoBERTa Large	7	0.196
	RoBERTa Base	8	0.178
Task 2	DeBERTa	4	0.339
	RoBERTa Large	6	0.390
	RoBERTa Base	5	0.277
Task 3	DeBERTa	4	0.323
	RoBERTa Large	9	0.374
	RoBERTa Base	3	0.308

Table 7: Results for simple regression.

Task	Model	Epoch	Macro-F1
Task 2	DeBERTa	11	0.293
	RoBERTa Large	5	0.360
	RoBERTa Base	8	0.274

Table 8: Results for Multi-class classification.

Task	Model	Epoch	RMSE mean
Task 2	DeBERTa	6	0.232
	RoBERTa Large	7	0.437
	RoBERTa Base	3	0.410

Table 9: Results for multi-output regression.

E Participant Results

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	CIMAT-NLP-GTO	0	**0.967**	**0.964**	**0.969**	**0.966**
2	UMUTeam	0	0.920	0.922	0.914	0.918
3	UNSL	1	0.913	0.912	0.920	0.913
4	UMUTeam	1	0.907	0.908	0.901	0.904
5	VICOM-nlp	2	0.880	0.878	0.885	0.879
6	VICOM-nlp	1	0.860	0.860	0.868	0.859
7	VICOM-nlp	0	0.853	0.850	0.850	0.850
8	CIMAT-NLP-GTO	1	0.847	0.868	0.866	0.847
9	PLN-CMM	0	0.827	0.856	0.849	0.827
10	CIMAT-NLP	1	0.820	0.836	0.837	0.820
11	BaseLine - DeBERTa	0	0.813	0.842	0.835	0.813
12	BaseLine - RoBERTa Large	1	0.813	0.823	0.827	0.813
13	CIMAT-NLP	0	0.807	0.844	0.831	0.806
14	NLP-UNED	0	0.760	0.792	0.783	0.760
15	UNSL	0	0.753	0.817	0.785	0.751
16	NLP-UNED	1	0.760	0.760	0.745	0.749
17	Xabi IXA	1	0.733	0.746	0.747	0.733
18	CIMAT-NLP-GTO	2	0.720	0.802	0.756	0.715
19	Xabi IXA	2	0.740	0.773	0.707	0.709
20	BaseLine - RoBERTa Base	2	0.700	0.783	0.736	0.694
21	Xabi IXA	0	0.693	0.688	0.691	0.689
22	I2C-UHU	0	0.653	0.762	0.696	0.641
23	UPM	0	0.453	0.719	0.523	0.349
24	UPM	1	0.453	0.719	0.523	0.349
25	UPM	2	0.453	0.719	0.523	0.349

Table 10: Binary Classification evaluation in Task 1.a. Ranking metric: Macro-F1.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	CIMAT-NLP-GTO	0	0.334	**0.018**	6	0.898	**0.863**
2	UNSL	1	0.433	0.045	8	0.857	0.776
3	CIMAT-NLP-GTO	1	0.379	0.065	6	0.898	0.761
4	VICOM-nlp	2	0.169	0.070	3	0.959	0.832
5	PLN-CMM	0	0.498	0.074	10	0.817	0.679
6	CIMAT-NLP	0	0.306	0.082	5	0.918	0.748
7	BaseLine - DeBERTa	0	0.310	0.083	5	0.918	0.751
8	VICOM-nlp	1	0.223	0.085	3	0.959	0.814
9	CIMAT-NLP	1	0.370	0.088	5	0.918	0.752
10	BaseLine - RoBERTa Large	1	**0.163**	0.099	**2**	**0.979**	0.792
11	UNSL	0	0.502	0.105	8	0.867	0.673
12	VICOM-nlp	0	0.226	0.111	3	0.959	0.794
13	UMUTeam	0	0.438	0.113	19	0.646	0.584
14	UMUTeam	1	0.441	0.116	19	0.646	0.573
15	NLP-UNED	0	0.268	0.118	3	0.959	0.738
16	CIMAT-NLP-GTO	2	0.435	0.119	6	0.898	0.676
17	BaseLine - RoBERTa Base	2	0.186	0.132	**2**	**0.979**	0.722
18	Xabi IXA	1	0.305	0.152	4	0.938	0.685
19	I2C-UHU	0	0.236	0.152	3	0.959	0.679
20	NLP-UNED	1	0.303	0.196	3	0.959	0.666
21	Xabi IXA	0	0.348	0.211	4	0.938	0.611
22	UPM	0	0.282	0.233	**2**	**0.979**	0.597
23	UPM	1	0.282	0.233	2	**0.979**	0.597
24	UPM	2	0.282	0.233	2	**0.979**	0.597
25	Xabi IXA	2	0.325	0.237	3	0.959	0.589

Table 11: Latency evaluation in Task 1.a. Ranking metric: ERDE30.

Rank	Team	Run	**RMSE**	Pearson_coefficient
1	BaseLine - RoBERTa Base	2	**0.178**	**0.906**
2	CIMAT-NLP-GTO	1	0.192	0.885
3	BaseLine - RoBERTa Large	1	0.196	0.890
4	CIMAT-NLP-GTO	2	0.200	0.864
5	CIMAT-NLP	1	0.229	0.810
6	BaseLine - DeBERTa	0	0.231	0.868
7	I2C-UHU	0	0.240	0.827
8	PLN-CMM	0	0.244	0.773
9	UMUTeam	1	0.255	0.811
10	UMUTeam	0	0.257	0.825
11	CIMAT-NLP	0	0.274	0.823
12	UPM	0	0.324	0.586
13	UPM	1	0.324	0.586
14	UPM	2	0.324	0.586
15	CIMAT-NLP-GTO	0	0.348	**0.906**
16	NLP-UNED	0	0.357	0.599
17	Xabi IXA	1	0.383	0.326
18	Xabi IXA	0	0.384	0.298
19	NLP-UNED	1	0.454	0.551
20	Xabi IXA	2	0.503	0.352

Table 12: Simple Regression evaluation in Task 1.b. Ranking metric: RMSE.

Rank	Team	Run	p@5	p@10	p@20	**p@30**
1	BaseLine - RoBERTa Large	1	0.800	0.800	**0.900**	**0.900**
2	CIMAT-NLP-GTO	0	**1.000**	**0.900**	**0.900**	0.867
3	BaseLine - DeBERTa	0	0.800	**0.900**	0.850	0.867
4	CIMAT-NLP-GTO	2	0.400	0.700	0.850	0.867
5	CIMAT-NLP	1	**1.000**	**0.900**	**0.900**	0.800
6	BaseLine - RoBERTa Base	2	**1.000**	0.800	0.850	0.800
7	CIMAT-NLP	0	0.600	0.600	0.700	0.767
8	PLN-CMM	0	0.600	0.700	0.800	0.733
9	NLP-UNED	1	0.600	0.500	0.700	0.700
10	NLP-UNED	0	0.800	0.600	0.650	0.700
11	UPM	0	**1.000**	0.800	0.750	0.700
12	UPM	1	**1.000**	0.800	0.750	0.700
13	UPM	2	**1.000**	0.800	0.750	0.700
14	UMUTeam	0	0.600	0.700	0.650	0.700
15	I2C-UHU	0	**1.000**	0.700	0.750	0.700
16	CIMAT-NLP-GTO	1	0.600	0.500	0.550	0.633
17	UMUTeam	1	**1.000**	0.700	0.650	0.600
18	Xabi IXA	2	0.600	0.600	0.700	0.533
19	Xabi IXA	1	0.400	0.700	0.650	0.533
20	Xabi IXA	0	0.600	0.600	0.600	0.467

Table 13: Ranking-based evaluation in Task 1.b at round 25. Ranking metric: p@30.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	UMUTeam	0	**0.738**	0.756	0.749	**0.737**
2	UNSL	1	**0.738**	**0.791**	**0.756**	0.733
3	UNSL	0	0.732	0.752	0.742	0.731
4	TextualTherapists	1	0.732	0.766	0.746	0.729
5	SINAI-SELA	0	0.725	0.775	0.742	0.720
6	UMUTeam	1	0.705	0.714	0.712	0.705
7	BaseLine - RoBERTa Large	1	0.698	0.759	0.718	0.690
8	SINAI-SELA	1	0.685	0.751	0.705	0.675
9	TextualTherapists	0	0.664	0.740	0.687	0.651
10	NLP-UNED	1	0.651	0.674	0.664	0.648
11	CIMAT-NLP	1	0.658	0.726	0.679	0.645
12	BaseLine - DeBERTa	0	0.664	0.788	0.691	0.642
13	CIMAT-NLP-GTO	0	0.651	0.732	0.674	0.635
14	VICOM-nlp	2	0.651	0.754	0.677	0.631
15	CIMAT-NLP-GTO	1	0.638	0.714	0.661	0.621
16	NLP-UNED	0	0.624	0.662	0.641	0.617
17	VICOM-nlp	1	0.638	0.735	0.663	0.616
18	GetitDone	0	0.611	0.628	0.622	0.609
19	BaseLine - RoBERTa Base	2	0.631	0.744	0.658	0.605
20	CIMAT-NLP-GTO	2	0.624	0.715	0.650	0.602
21	Ana Laura Lezama Sánchez	0	0.577	0.576	0.577	0.576
22	VICOM-nlp	0	0.591	0.693	0.619	0.559
23	NLPUTB	0	0.604	0.619	0.579	0.554
24	NLPUTB	1	0.604	0.619	0.579	0.554
25	NLPUTB	2	0.604	0.619	0.579	0.554
26	TextualTherapists	2	0.577	0.698	0.608	0.537
27	PLN-CMM	0	0.517	0.697	0.554	0.434
28	DepNLP UC3M GURUDASI	0	0.483	0.734	0.525	0.366
29	DepNLP UC3M GURUDASI	1	0.483	0.734	0.525	0.366
30	DepNLP UC3M GURUDASI	2	0.483	0.734	0.525	0.366
31	SPIN	1	0.470	0.731	0.512	0.340
32	CIMAT-NLP	0	0.463	0.563	0.505	0.337
33	SPIN	0	0.463	0.730	0.506	0.327

Table 14: Binary classification evaluation in Task 2.a. Ranking metric: Macro-F1.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	SINAI-SELA	0	0.395	**0.140**	4.000	0.951	**0.720**
2	UNSL	1	0.567	0.148	14.000	0.791	0.609
3	BaseLine - DeBERTa	0	0.303	0.153	**2.000**	**0.984**	0.719
4	BaseLine - RoBERTa Large	1	0.290	0.159	4.000	0.951	0.704
5	SINAI-SELA	1	0.389	0.159	4.000	0.951	0.696
6	TextualTherapists	1	0.421	0.161	7.000	0.903	0.682
7	TextualTherapists	0	0.342	0.168	3.000	0.967	0.696
8	VICOM-nlp	2	**0.275**	0.173	**2.000**	**0.984**	0.706
9	CIMAT-NLP-GTO	0	0.423	0.175	5.000	0.935	0.665
10	BaseLine - RoBERTa Base	2	0.342	0.176	4.000	0.951	0.671
11	VICOM-nlp	1	0.281	0.183	**2.000**	**0.984**	0.695
12	CIMAT-NLP	1	0.290	0.187	3.000	0.967	0.689
13	CIMAT-NLP-GTO	1	0.412	0.187	4.000	0.951	0.666
14	UNSL	0	0.551	0.188	14.000	0.791	0.591
15	CIMAT-NLP-GTO	2	0.414	0.199	4.000	0.951	0.662
16	VICOM-nlp	0	0.289	0.201	**2.000**	**0.984**	0.666
17	TextualTherapists	2	0.330	0.205	**2.000**	**0.984**	0.663
18	NLP-UNED	1	0.411	0.207	6.000	0.919	0.624
19	NLP-UNED	0	0.404	0.212	5.000	0.935	0.627
20	PLN-CMM	0	0.286	0.224	**2.000**	**0.984**	0.640
21	DepNLP UC3M GURUDASI	0	0.349	0.236	3.000	0.967	0.618
22	DepNLP UC3M GURUDASI	1	0.349	0.236	3.000	0.967	0.618
23	DepNLP UC3M GURUDASI	2	0.349	0.236	3.000	0.967	0.618
24	GetitDone	0	0.302	0.240	**2.000**	**0.984**	0.627
25	SPIN	1	0.402	0.242	3.000	0.967	0.612
26	SPIN	0	0.431	0.245	3.000	0.967	0.609
27	CIMAT-NLP	0	0.315	0.249	**2.000**	**0.984**	0.616
28	NLPUTB	0	0.362	0.356	**2.000**	**0.984**	0.397
29	NLPUTB	1	0.362	0.356	**2.000**	**0.984**	0.397
30	NLPUTB	2	0.362	0.356	**2.000**	**0.984**	0.397
31	UMUTeam	0	0.548	0.358	30.000	0.560	0.421
32	UMUTeam	1	0.548	0.371	30.000	0.560	0.398
33	Ana Laura Lezama Sánchez	0	0.561	0.561	101.000	0.074	0.041

Table 15: Latency evaluation in Task 2.a. Ranking metric: ERDE30.

Rank	Team	Run	**RMSE**	Pearson_coefficient
1	BaseLine - RoBERTa Base	2	**0.277**	**0.770**
2	CIMAT-NLP-GTO	1	0.292	0.645
3	CIMAT-NLP-GTO	2	0.294	0.630
4	PLN-CMM	0	0.309	0.642
5	UMUTeam	1	0.325	0.522
6	UMUTeam	0	0.333	0.484
7	CIMAT-NLP	1	0.335	0.661
8	BaseLine - DeBERTa	0	0.339	0.683
9	CIMAT-NLP-GTO	0	0.367	0.632
10	NLPUTB	0	0.381	0.318
11	NLPUTB	1	0.381	0.318
12	NLPUTB	2	0.381	0.318
13	BaseLine - RoBERTa Large	1	0.390	0.503
14	NLP-UNED	0	0.401	0.317
15	DepNLP UC3M GURUDASI	0	0.405	0.196
16	DepNLP UC3M GURUDASI	1	0.405	0.196
17	DepNLP UC3M GURUDASI	2	0.405	0.196
18	NLP-UNED	1	0.406	0.358
19	CIMAT-NLP	0	0.540	0.054

Table 16: Simple Regression evaluation in Task 2.b. Ranking metric: RMSE.

Rank	Team	Run	p@5	p@10	p@20	**p@30**
1	PLN-CMM	0	**0.800**	**0.800**	**0.700**	**0.600**
2	BaseLine - RoBERTa Large	1	0.400	0.500	0.550	0.567
3	CIMAT-NLP-GTO	0	0.600	0.600	0.500	0.567
4	BaseLine - DeBERTa	0	**0.800**	0.600	0.550	0.567
5	CIMAT-NLP-GTO	1	0.600	0.500	0.550	0.567
6	BaseLine - RoBERTa Base	2	0.600	**0.800**	**0.700**	0.567
7	CIMAT-NLP	1	0.600	0.600	0.550	0.533
8	CIMAT-NLP-GTO	2	0.600	0.400	0.450	0.533
9	NLP-UNED	1	0.200	0.400	0.350	0.367
10	UMUTeam	1	0.400	0.500	0.350	0.333
11	CIMAT-NLP	0	0.000	0.000	0.250	0.300
12	UMUTeam	0	0.400	0.200	0.350	0.300
13	DepNLP UC3M GURUDASI	0	0.400	0.300	0.350	0.267
14	DepNLP UC3M GURUDASI	1	0.400	0.300	0.350	0.267
15	DepNLP UC3M GURUDASI	2	0.400	0.300	0.350	0.267
16	NLP-UNED	0	0.200	0.400	0.350	0.267
17	NLPUTB	0	0.000	0.000	0.000	0.000
18	NLPUTB	1	0.000	0.000	0.000	0.000
19	NLPUTB	2	0.000	0.000	0.000	0.000

Table 17: Ranking-based evaluation in Task 2.b at rank 25. Ranking metric: p@30.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	BaseLine - RoBERTa Large	1	0.483	0.389	0.378	**0.360**
2	NLP-UNED	1	**0.490**	0.366	**0.389**	0.358
3	NLP-UNED	0	0.450	0.362	0.375	0.339
4	BaseLine - DeBERTa	0	0.456	0.395	0.344	0.293
5	PLN-CMM	0	0.383	0.329	0.327	0.288
6	BaseLine - RoBERTa Base	2	0.356	0.380	0.335	0.274
7	I2C-UHU	0	0.315	0.307	0.253	0.232
8	DepNLP UC3M GURUDASI	0	0.322	0.362	0.315	0.227
9	DepNLP UC3M GURUDASI	1	0.322	0.362	0.315	0.227
10	DepNLP UC3M GURUDASI	2	0.322	0.362	0.315	0.227
11	SPIN	1	0.262	0.412	0.343	0.219
12	SPIN	0	0.255	0.384	0.297	0.190
13	SPIN	2	0.248	**0.434**	0.292	0.161

Table 18: Multiclass classification evaluation in Task 2.c. Ranking metric: Macro-F1.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	BaseLine - DeBERTa	0	0.330	**0.190**	**2.000**	**0.984**	**0.695**
2	I2C-UHU	0	**0.272**	0.198	**2.000**	**0.984**	0.670
3	NLP-UNED	1	0.412	0.203	5.000	0.935	0.638
4	BaseLine - RoBERTa Base	2	0.307	0.206	**2.000**	**0.984**	0.659
5	NLP-UNED	0	0.408	0.211	5.000	0.935	0.627
6	DepNLP UC3M GURUDASI	0	0.383	0.228	3.000	0.967	0.632
7	DepNLP UC3M GURUDASI	1	0.383	0.228	3.000	0.967	0.632
8	DepNLP UC3M GURUDASI	2	0.383	0.228	3.000	0.967	0.632
9	BaseLine - RoBERTa Large	1	0.283	0.232	**2.000**	**0.984**	0.652
10	PLN-CMM	0	0.348	0.232	**2.000**	**0.984**	0.645
11	SPIN	1	0.402	0.242	3.000	0.967	0.612
12	SPIN	0	0.431	0.245	3.000	0.967	0.609
13	SPIN	2	0.431	0.245	3.000	0.967	0.609

Table 19: Latency evaluation in Task 2.c. Ranking metric: ERDE30.

Rank	Team	Run	**RMSE_mean**	RMSE_sf	RMSE_sa	RMSE_so	RMSE_c	Pearson_mean	Pearson_sf	Pearson_sa	Pearson_so	Pearson_c
1	BaseLine - DeBERTa	0	**0.232**	**0.246**	0.250	**0.125**	**0.306**	**0.484**	**0.661**	0.295	**0.260**	**0.721**
2	DepNLP UC3M GURUDASI	0	0.250	0.272	0.228	0.129	0.371	0.131	0.207	0.018	0.057	0.240
3	DepNLP UC3M GURUDASI	1	0.250	0.272	0.228	0.129	0.371	0.131	0.207	0.018	0.057	0.240
4	DepNLP UC3M GURUDASI	2	0.250	0.272	0.228	0.129	0.371	0.131	0.207	0.018	0.057	0.240
5	PLN-CMM	0	0.349	0.328	**0.210**	0.391	0.469	-0.052	0.051	0.394	-0.153	-0.498
6	BaseLine - RoBERTa Base	2	0.410	0.547	0.272	0.235	0.585	-0.145	-0.496	0.355	0.185	-0.624
7	BaseLine - RoBERTa Large	1	0.437	0.682	0.312	0.158	0.598	-0.209	-0.678	**0.890**	0.059	-0.306

Table 20: Multi-output Regression evaluation for Task 2.d. Metric ranking: RMSE_mean.

Rank	Team	Run	p@5	p@10	p@20	**p@30**	p@5
1	DepNLP UC3M GURUDASI	0	**0.350**	**0.400**	**0.375**	**0.350**	**0.400**
2	DepNLP UC3M GURUDASI	1	**0.350**	**0.400**	**0.375**	**0.350**	**0.400**
3	DepNLP UC3M GURUDASI	2	**0.350**	**0.400**	**0.375**	**0.350**	**0.400**
4	BaseLine - RoBERTa Large	1	**0.350**	0.275	0.263	0.275	0.350
5	BaseLine - DeBERTa	0	0.250	0.300	0.338	**0.350**	0.250
6	BaseLine - RoBERTa Base	2	0.300	0.300	0.225	0.192	0.250
7	PLN-CMM	0	0.250	0.200	0.200	0.175	0.200

Table 21: Ranking-based evaluation for Task 2.d at round 25. Ranking metric: p@30.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	**Macro-F1**
1	CIMAT-NLP-GTO	2	**0.773**	0.780	**0.729**	**0.740**
2	BaseLine - DeBERTa	0	0.760	**0.840**	0.688	0.693
3	NLP-UNED	1	0.680	0.657	0.647	0.650
4	BaseLine - RoBERTa Large	1	0.720	0.795	0.638	0.630
5	CIMAT-NLP	0	0.673	0.654	0.614	0.614
6	NLP-UNED	0	0.640	0.609	0.594	0.595
7	CIMAT-NLP-GTO	0	0.633	0.602	0.592	0.593
8	BaseLine - RoBERTa Base	2	0.680	0.755	0.586	0.553
9	CIMAT-NLP-GTO	1	0.653	0.671	0.557	0.516
10	CIMAT-NLP	1	0.627	0.600	0.519	0.444
11	UPM	0	0.627	0.812	0.509	0.402
12	UPM	1	0.627	0.812	0.509	0.402
13	UPM	2	0.627	0.812	0.509	0.402

Table 22: Binary classification evaluation for Task 3.a.

Rank	Team	Run	ERDE5	**ERDE30**	latencyTP	speed	latency-weightedF1
1	BaseLine - DeBERTa	0	0.347	**0.165**	4.000	0.954	0.798
2	BaseLine - RoBERTa Large	1	0.324	0.179	**2.000**	**0.985**	**0.800**
3	CIMAT-NLP-GTO	2	0.691	0.188	7.000	0.908	0.757
4	BaseLine - RoBERTa Base	2	**0.309**	0.210	**2.000**	**0.985**	0.779
5	UPM	0	0.341	0.231	**2.000**	**0.985**	0.757
6	UPM	1	0.341	0.231	**2.000**	**0.985**	0.757
7	UPM	2	0.341	0.231	**2.000**	**0.985**	0.757
8	CIMAT-NLP-GTO	1	0.753	0.232	7.000	0.908	0.703
9	CIMAT-NLP	1	0.839	0.247	14.000	0.802	0.612
10	CIMAT-NLP	0	0.769	0.250	14.000	0.802	0.614
11	CIMAT-NLP-GTO	0	0.710	0.283	7.000	0.908	0.654
12	NLP-UNED	1	0.632	0.285	8.000	0.893	0.672
13	NLP-UNED	0	0.652	0.310	8.000	0.893	0.652

Table 23: Latency evaluation in Task 3.a. Ranking metric: ERDE30.

Rank	Team	Run	**RMSE**	Pearson_coefficient
1	BaseLine - RoBERTa Base	2	**0.308**	**0.693**
2	BaseLine - DeBERTa	0	0.323	0.682
3	CIMAT-NLP-GTO	1	0.329	0.497
4	CIMAT-NLP	1	0.332	0.468
5	CIMAT-NLP-GTO	2	0.348	0.576
6	CIMAT-NLP-GTO	0	0.367	0.385
7	BaseLine - RoBERTa Large	1	0.374	-0.092
8	UPM	0	0.435	0.191
9	UPM	1	0.435	0.191
10	UPM	2	0.435	0.191
11	CIMAT-NLP	0	0.472	0.324
12	NLP-UNED	0	0.481	0.172
13	NLP-UNED	1	0.482	0.243

Table 24: Simple Regression evaluation for Task 3.b. Metric ranking: RMSE.

Rank	Team	Run	p@5	p@10	p@20	**p@30**
1	BaseLine - DeBERTa	0	0.800	0.600	0.700	**0.767**
2	BaseLine - RoBERTa Base	2	0.800	0.700	0.750	0.700
3	CIMAT-NLP-GTO	2	**1.000**	0.800	**0.750**	0.667
4	UPM	0	0.600	0.500	0.600	0.633
5	UPM	1	0.600	0.500	0.600	0.633
6	UPM	2	0.600	0.500	0.600	0.633
7	CIMAT-NLP-GTO	1	**1.000**	**0.900**	0.650	0.533
8	CIMAT-NLP	0	0.600	0.500	0.400	0.500
9	CIMAT-NLP-GTO	0	**1.000**	0.800	0.500	0.467
10	CIMAT-NLP	1	0.600	0.500	0.450	0.467
11	NLP-UNED	0	0.200	0.300	0.450	0.400
12	NLP-UNED	1	0.200	0.300	0.350	0.333
13	BaseLine - RoBERTa Large	1	0.200	0.100	0.350	0.300

Table 25: Ranking-based evaluation for Task 3.b. at round 25. Ranking metric: p@30.