UMUTeam at MentalRiskES2023@IberLEF: Transformer and Ensemble Learning Models for Early Detection of Eating Disorders and Depression

Ronghao Pan[1,*,†], José Antonio García-Díaz[1,†] and Rafael Valencia-García[1,†]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

This paper describes the participation of the UMUTeam in MentalRiskEs@IberLEF 2023. We have addressed all subtasks of Task 1 and 2 first subtasks of Task 2 (Task 2.a and Task 2.b). For this purpose, a fine-tuning approach of pre-trained monolingual and multilingual models has been proposed for the binary classification problem. An ensemble learning approach was also tested. This approach concatenates the pooler output of the pre-trained model with the last hidden state of a pre-trained model for emotion identification. As for the regression problem, instead of building a model from scratch, the obtained classification models were used and a softmax transformation was applied to the output to obtain the probability of suffering. In Task 1.a, our team placed second with a macro F1 score of 0.918 in the decision-based evaluation and 13th in the latency-based evaluation. In Task 1.b, our team placed 9th with an RMSE of 0.255 in the regression-based evaluation and 14th in the ranking-based evaluation. For Task 2.a, our team ranked first with a macro F1 score of 0.737 in the classification-based evaluation and 31st in the latency-based evaluation. In the simple regression problem of Task 2 (Task 2.b) our team achieved the fifth-best result with an RMSE of 0.325 in the regression-based evaluation and 10th place in the ranking-based evaluation.

Keywords

Natural Language Processing, Transformers, Large Language Model, Ensemble learning, Text classification

1. Introduction

Recently, there has been a significant increase in the number of people suffering from mental disorders such as anxiety, depression, and eating disorders. According to a recent World Health Organization report, 1 in 8 people suffer from a mental disorder. Therefore, early detection is a key effective intervention to prevent these problems.

Social networks have become a useful source of data for the early detection of mental disorders and health problems due to the large amount of content that is posted on them on a daily basis. As a result, there is a growing interest in detecting and identifying mental disorders in social media streams, and several collaborative tasks have emerged that aim at the early detection of different types of mental risks, such as eating disorders, dysthymia, anxiety, depression, and

IberLEF 2023, September 2023, Jaén, Spain

ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

suicidal ideation, among others. One of the most prominent is eRisk (Early-Risk Identification Task) [1], which is hosted by relevant evaluation campaigns such as CLEF (Cross-Lingual Evaluation Forum). However, all this work has focused mainly on English.

As a result, the MentalRiskEs [2] task was created, which consists of a novel task of early detection of mental disorders risks in Spanish comments from Telegram users. The MentalRiskEs@IberLEF 2023 task focuses on the early detection of signs of eating disorders, depression, and unknown disorders in order to evaluate the robustness of approaches to new disorders that are not known a priori. In this edition, three tasks have been proposed:

- **Task 1. Eating disorders detection**: It is a task that focuses on detecting whether the user suffers from anorexia or bulimia. This task is divided into 2 subtasks, one that considers it as a binary classification problem and the other as a regression problem on the probability that the user suffers from anorexia or bulimia.
- **Task 2. Depression detection**: This task aims at the early detection of depressive symptoms. For this purpose, the task is divided into 4 subtasks: (1) treating this task as a binary classification problem, (2) treating this task as a regression problem, with the aim of indicating the probability that the user is suffering from depression, (3) considering this task as a multi-class classification problem, involving four different classes (suffer+against, suffer+for, suffer+other and control) indicating different states of the user, and (4) considering the multi-class classification problem as a regression problem.
- **Task 3. Non-defined disorder detection**: This task focuses on the detection of undefined disorders that are related to the above-mentioned disorders (eating disorders and depression). For this purpose, this task is divided into two subtasks, one as a binary classification problem (suffer, control) and the other as a regression problem. In this case, the organizers recommend using the models obtained in the previous tasks to identify another disorder unknown to the participant.

The tasks had to be solved as an online problem, i.e., the participants had to be able to identify a potential risk as early as possible in a continuous stream of data. Therefore, the participants' system was tasked with reading and processing messages from multiple users sequentially in order to generate a response to obtain the next posts.

This paper presents UMUTeam's contribution to Tasks 1.a, 1.b, 2.a, and 2.b. For these tasks, we have used different approaches such as fine-tuning different pre-trained transformer-based models and ensemble learning to combine these models with a pre-trained emotion detection model to address the binary classification problem. In the case of regression problems, instead of training a model from scratch, binary classification models have been used to obtain a probability that a user suffers from a mental disorder problem. The rest of the paper is organized as follows. Section 2 presents the task and the dataset provided. It also describes the methodology of our proposed system for addressing each task. Secondly, Section 3 shows the results obtained, and a discussion of them is presented. Finally, Section 4 concludes the paper with some findings and perspectives for future work.

2. Task description and Methodology

In this section, we present the different MentalRiskEs tasks that we have participated in, aimed at the early detection of symptoms of depression and eating disorders, as well as the dataset provided to perform them. In addition, we describe the methodology of our proposed system to address each task.

2.1. Task 1. Eating disorders detection

This task consists in the early detection of eating disorders through different Spanish comments of Telegram users. It is divided into two subtasks, one that considers it as a binary classification problem (suffer, control) and another that considers it as a regression problem. The dataset provided consists of a set of comments from 175 users, of whom 74 suffer from bulimia or anorexia, with a total of 2532 comments, and 101 users who do not suffer from any eating disorder, with a total of 3399 comments.

The task is approached from two different perspectives: as a binary decision problem and as a regression decision problem. As a binary classification problem, user comments should be labeled as positive (label 1, i.e., suffers detected) or negative (label 0, i.e., control detected) depending on the user's state. Thus, in the dataset, we have a total of 2532 comments indicating that the user suffers from a bulimia or anorexia problem and 3399 that it is not related to bulimia and anorexia. As for the regression problem, each user in the dataset has a probability indicating whether he/she suffers from an eating disorder or not. A value of 0 means 100% negative, and a value of 1 would be 100% positive.

To address this task, we followed a supervised learning approach. To train our model at the sentence level, we selected a custom split from the provided training data for validation. The custom validation split is created using stratified sampling to maintain the balance between labels. The final training and validation set is shown in Table 1.

Table 1
The distribution of the training and validation split for Task 1.a (messages level).

	Suffer	Control	Total
Training	2 025	2 719	4 744
Validation	507	680	1 187

2.2. Task 2. Depression detection

This task aims at the early detection of depression through different Spanish comments of Telegram users and is approached from four different perspectives: as a binary classification model (Task 2.a), as a regression problem for binary classification (Task 2.b), as a multi-class classification (Task 2.c) and as a regression problem for multi-class classification (Task 2.d). In this case, our team was involved in the first two subtasks, consisting of the binary classification problem and a simple regression indicating a probability of suffering from depression. The data provided consists of a set of comments from 175 users, of which 94 users suffer from depression

with a total of 3113 comments, and 81 users do not suffer from depression with a total of 3135 comments. To address this task, we followed a supervised learning approach. To train our model at the sentence level, we selected a custom split for validation. The custom validation split is created using stratified sampling, in order to keep the balance among the labels. The final set of training and validation is shown in Table 2.

Table 2
The distribution of the training and validation split for Task 2.a (messages level).

	Suffer	Control	Total
Training	2 490	2 508	4 998
Validation	623	627	1 250

2.3. System and methods for Task 1 and Task 2

To address this task as a binary classification problem, we followed a supervised learning approach, which is based on fine-tuning different pre-trained models in Spanish and selecting the one with the best result. We have also evaluated the addition of emotion to the model using an ensemble learning approach that combines the pooler output of the pre-trained transformer-based models with the last hidden state of a pre-trained emotion detection model. The pooler output is the last layer of the hidden state of the first token of the sequence after further processing through the layers used for the auxiliary pre-training task, and the last hidden state is a sequence of hidden states at the output of the last layer of the model. To extract the emotion features, we used a pre-trained language model based on the Transformer architecture called *pysentimiento*[3], which is a BETO-based model fine-tuned on an annotated Twitter corpus on emotions.

Figure 1 shows the pipeline used for this task, which can be described as follows. First, the dataset is processed by removing all mentions and hashtags and extracting all emoji features using the *emoji* library. Second, after the data pre-processing, to create a binary classification model, we tested the fine-tuning of different Spanish pre-trained models such as BETO and MarIA. For this purpose, we first tokenized the text through a corresponding tokenization of each model, and then we performed the fine-tuning process, which consists of adapting the model to a classification task, and finally adding a final layer of sequence classification is added for the classification task. The models were trained with 10 epochs, a learning rate of 2e-5, and a weight decay of 0.01. The Spanish pre-trained models were evaluated as follows: (1) BETO [4], (2) ALBETO [5], (3) DistilBETO [5], and (4) MarIA [6]. We also evaluate XLM-RoBERTa as a multilingual transformer [7].

Finally, for the ensemble learning approach, we concatenated the pooler output of one of the pre-trained models with the last hidden state of the pre-trained emotion identification model (*pysentimiento*) with 20 epochs and 2e-6 of learning rate.

For the regression problem (Task 1.b and Task 2.b), we used the binary classification models (Task 1.a and Task 2.b), adding a softmax transformation to obtain the probability of each label. In this way, the probability that a user suffers from anorexia or bulimia is the average of all the probabilities of the predictions of the label *suffer* on the user's comments.

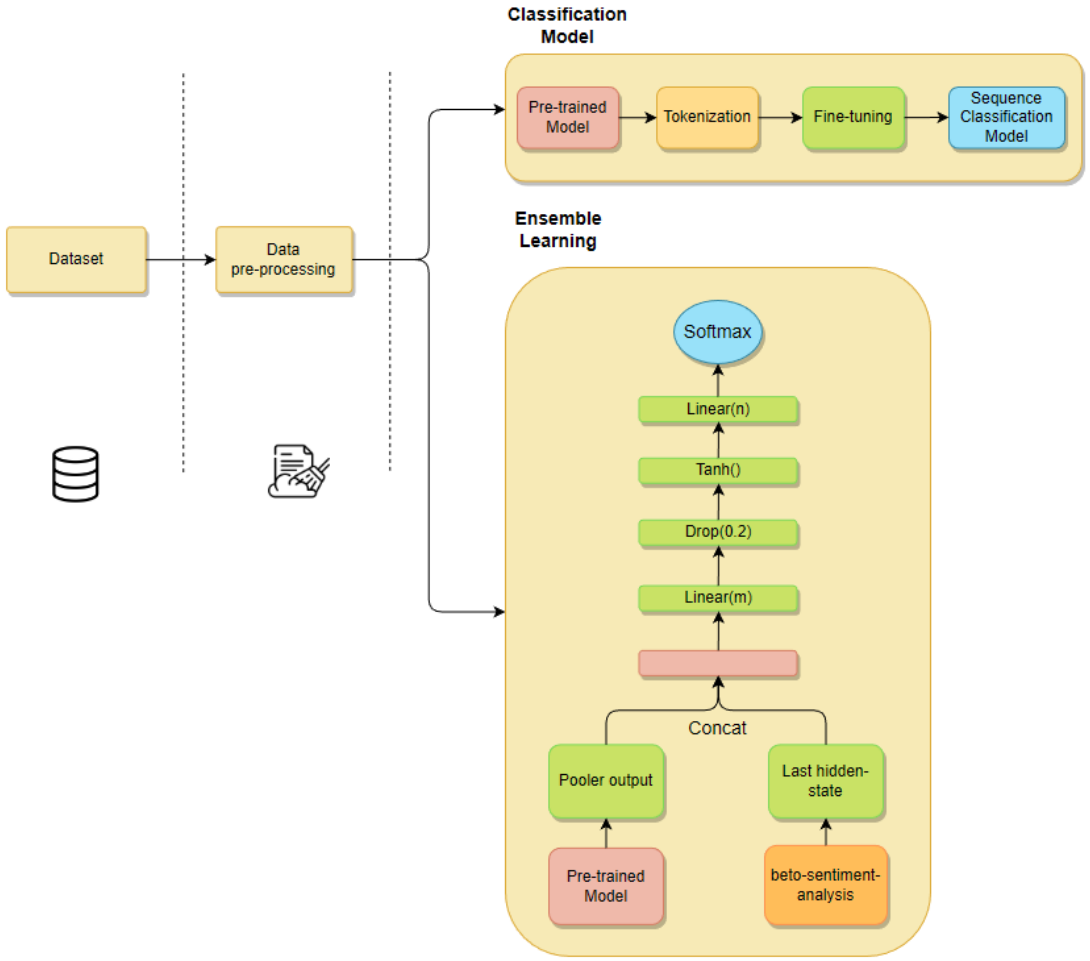

Figure 1: System architecture of Task 1 and Task 2.

3. Results

This section describes the systems submitted by our team in each run and shows the results obtained in each task.

3.1. Task 1. Eating disorders detection

For Task 1.a, different pre-training models were tested and the results of the validation split are shown in Table 3. On the one hand, it can be observed that the MarIA model has obtained the best performance compared to other models and, on the other hand, it has been detected that monolingual models such as BETO and MarIA, have a better performance than complex multilingual models such as XLM-R. Thus, for the ensemble learning approach, a model has been introduced that concatenates the pooler output of the MarIA with the last hidden state of

pysentimiento and has an improvement of 0.86% over MarIA in the macro F1 metric.

Table 3
Benchmark of the different pre-trained and ensemble learning models with validation splits in Task 1. For each model, the macro precision (M-P), macro recall (M-R), and macro (M-F1) are reported.

Model	M-P	M-R	M-F1
BETO	0.7357	0.7249	0.7278
MarIA	0.7357	0.7299	**0.7319**
XLM-R	0.7095	0.7111	0.7102
ALBETO	0.6935	0.6931	0.6933
DistilBETO	0.7116	0.7113	0.7114
MarIA + pysentimiento	0.7461	0.7380	**0.7405**

However, the models obtained are at the sentence level, so predicting whether a user has an eating disorder requires taking into account predictions from the user's previous comments, since a user's comment about bulimia or anorexia does not ensure that he or she actually suffers from that disorder. Therefore, to determine whether a user is suffering from bulimia or anorexia, the set of user messages is run through our system and decided by the most repeated label. For the early detection approach, it has been tested to set a limit of comments to be taken into account, for example, if the limit is 5 and the first 4 user comments are related to bulimia or anorexia then the system will make the decision to suffer. After testing several thresholds, we found that the system works better when all user comments are processed. So our system uses a conservative early detection strategy, which is to make a decision after processing all user comments.

For the binary classification problem (Task 1.a), we submitted two runs for this task. The runs have the same structure but differ in minor aspects of configuration.

- **Run 0**: This run consists of using the best binary classification model at the sentence level and running this model with pre-processing on the set of comments used.
- **Run 1**: This run has the same structure as Run 0, but uses the ensemble learning model.

From the results reported by the organizers, we have extracted our scores, which are shown in Table 4. In the decision-based evaluation, the best macro F1 obtained was Run 0 (0.914), which is the second highest among all participant submissions (a total of 25 submissions were reported). In Run 1, it obtained a macro F1 of 0.904, which is the fourth highest among all submissions. In terms of latency-based evaluation, our team has achieved 13th place with Run 0 and 14th place with Run 1, with both runs using the conservative early detection strategy. In terms of carbon emissions, our approach has a mean duration of 11.51009418 with a deviation of 6.123578938 and a mean emission of 7.01E-08 with a deviation of 3.29E-07 in the best run (Run 0). In this case, our mean duration is in the middle of the rankings, but our mean emission is one of the lowest.

For the error analysis, we carried out a study of the confusion matrix on the best run (Run 0), as it gives us a more detailed picture of the system's behavior when performing a classification.

Table 4
Results of UMUTeam for Task 1.a in decision-based and latency-based evaluation. For each run, the macro precision (M-P), macro recall (M-R), and macro (M-F1) are reported for decision-based evaluation, and ERDE5, ERDE30, latencyTP, speed, latency-weightedF1 for latency-based evaluation.

	Accuracy	M-P	M-R	M-F1	ERDE5	ERDE30	latencyTP	speed	latency weightedF1
Run 0	0.92	0.922	0.914	0.918	0.438	0.113	0.19	0.646	0.584
Run 1	0.907	0.908	0.901	0.904	0.441	0.116	0.19	0.646	0.573

In this case, we have used a normalized confusion matrix, where each row represents an instance of the actual class, while a column represents an instance of the predicted class. Therefore, by using a normalized confusion matrix, the values of the diagonal elements represent the degree of correctly predicted classes. The confusion is expressed by the misclassified non-diagonal elements, as they are confused with another class. Figure 2 shows the confusion matrix in Run 0 of our system. It can be observed that it has good accuracy in identifying people without eating disorders with a probability of 95.35%, but it makes errors of 12.50% in classifying people with eating disorders.

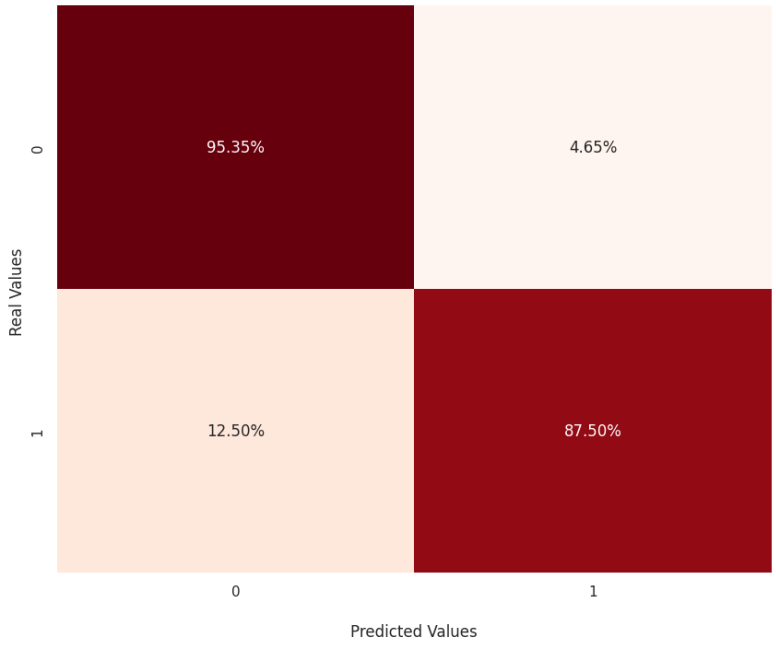

Figure 2: Confusion matrix on the best run (Run 0) of Task 1.a.

With respect to the regression problem (Task 2.b), instead of training a model from scratch, we have reused binary classification models to obtain the probability that a comment was related to eating disorders. To do this, we added a softmax layer to the ensemble learning and fine-tuned MarIA to obtain the probability of each label. To obtain the probability that a user

suffers from eating disorders, we used the conservative strategy, which consists of processing all the user's comments and averaging the probabilities of the comments of the suffering type.

For Task 1.b, we presented two runs. The runs have the same structure but with different binary classification models.

- **Run 0**: This run consists of using the fine-tuned MarIA and running this model where the set of comments used has a pre-processing.
- **Run 1**: This run shares the same structure as Run 0, although it uses the ensemble learning model.

Table 5
Results of UMUTeam for Task 1.b in regression-based and ranking-based evaluation. For each run, RMSE and Pearson coefficient are reported for decision-based evaluation and p@5, p@10, p@20, and p@30 for ranking-based evaluation.

	RMSE	Pearson coefficient	p@5	p@10	p@20	p@30
Run 0	0.257	0.825	0.60	0.70	0.65	0.70
Run 1	0.255	0.811	1.00	0.70	0.65	0.60

As can be seen from the results obtained in the regression problem (see Table 5, Run 1 has obtained a better result (0.255 in RSME) than Run 0 (0.257 in RSME) in the regression-based evaluation, reaching the ninth and tenth best results of all submissions (a total of 20 submissions in total were reported). As for the ranking-based evaluation, the best result was obtained with Run 1 using the ensemble learning model, which achieved the 14th-best result out of a total of 20 submissions. The main reason why we did not perform so well in the ranking-based evaluation is that we have used the conservative strategy for early detection, so our system needs to process all user comments in order to make a decision. As for the carbon emission of our best execution (Run 1), it has a mean duration of 11.26519856 with a deviation of 6.281654428 and a mean emission of 6.86E-08 with a deviation of 3.25E-07.

To perform error analysis on the regression model, we have created a graph comparing the predicted values with the actual values. Figure 3 shows the graph obtained by comparing the values predicted by our best run (Run 1) with the actual values. It can be observed that our model has the same trends as the actual values, and it can be seen that in many cases when our model identifies a user with a value higher than 0.8, their actual value is 1.

3.2. Task 2. Depression detection

As explained in Section 2, we have used the same approach as in Task 1 to solve Task 2.a and 2.b, which aim to detect the sign of depression in the user. Therefore, we used the same experimental setup and runs as in Task 1 (see section 3.1). Table 6 shows the results obtained with the classification models in the validation set, and it can be seen that in this case, the BETO obtained the best result with macro F1 of 0.6651. It can also be seen that the monolingual

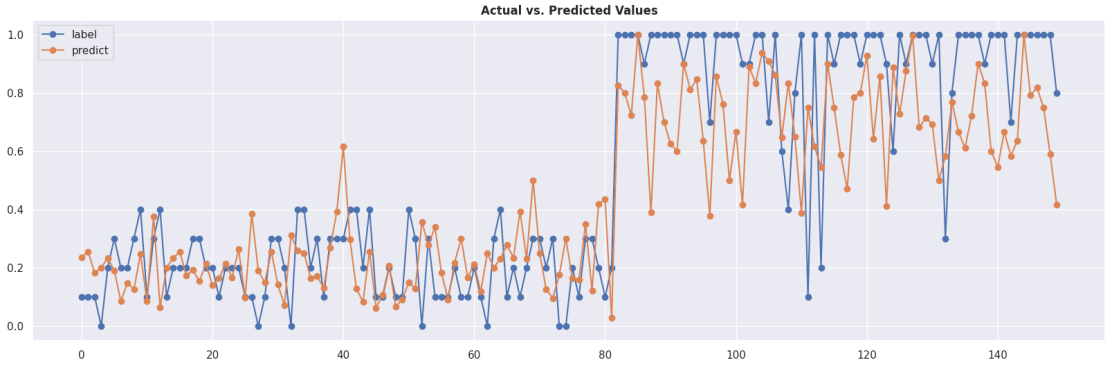

Figure 3: Comparison chart between actual and predicted values for Task 1.b.

models (BETO, MarIA, ALBETO and DistilBETO) obtained better results than the multilingual model. Regarding the ensemble learning model, BETO was combined with *pysentimiento*, and an improvement of 0.044% was obtained over the best fine-tuned model for binary classification (BETO).

Table 6
Benchmark of the different pre-trained and ensemble learning models with validation splits in Task 2. For each model, the macro precision (M-P), macro recall (M-R), and macro (M-F1) are reported.

	M-P	M-R	M-F1
BETO	0.6669	0.6657	**0.6651**
MarIA	0.6289	0.6288	0.6287
XLM-R	0.6426	0.6425	0.6423
ALBETO	0.6003	0.5999	0.5995
DistilBETO	0.6442	0.6441	0.6439
BETO + pysentimiento	0.6698	0.6696	**0.6695**

Table 7 shows the results obtained in the decision-based evaluation with the runs presented in Task 2.a. The value of the macro F1 score obtained by our Run 0 (0.737) is the first highest among all submissions by participants (33 submissions were reported in total). In the latency-based evaluation, Run 0 performed better than Run 1, with an ERDE30 of 0.358. However, by using a conservative strategy for early detection, we obtained a worse ranking result (31st) than the decision-based one, as our system needs to process all user comments to make a decision. In addition, Run 0 has one of the lowest mean emissions, with a value of 5.52E-08 and a deviation of 5.44E-07. However, the mean duration is 19.49339786 with a deviation of 19.88332675.

Figure 4 shows the confusion matrix for Run 0 of our system. It can be observed that our model is biased toward identification 1 (users with depression) with 94.12%. However, it fails in many cases in the identification of 0 (users without depression) with a probability of 39.51%.

In Task 2.b, we have presented two runs as in Task 1.b (see Section 3.1), both with the same structure and decision strategy, but with different models. For Run 0, we have used the fine-

Table 7
Results of UMUTeam for Task 2.a in decision-based and latency-based evaluation. For each run, the macro precision (M-P), macro recall (M-R), and macro (M-F1) are reported for decision-based evaluation, and ERDE5, ERDE30, latencyTP, speed, latency-weightedF1 for latency-based evaluation.

	Accuracy	M-P	M-R	M-F1	ERDE5	ERDE30	latencyTP	speed	latency weightedF1
Run 0	0.738	0.756	0.749	0.737	0.548	0.358	30.000	0.560	0.421
Run 1	0.705	0.714	0.712	0.705	0.548	0.371	30.000	0.560	0.398

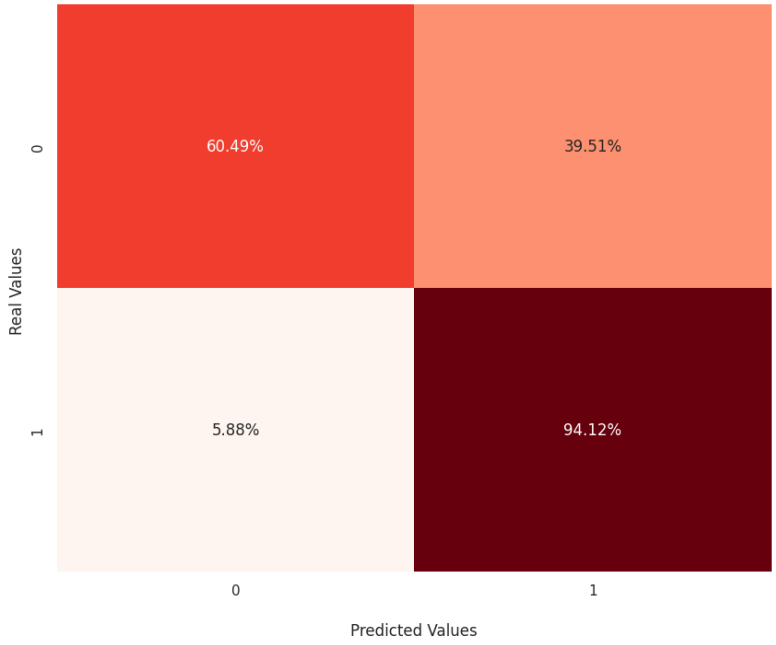

Figure 4: Confusion matrix on the best run (Run 0) of Task 2.a.

tuned BETO with a softmax transformation at the end to obtain the probability. On the other hand, for Run 1, the ensemble learning model was used to obtain the probability of each label.

The results obtained in the regression-based and ranking-based evaluation of the two runs presented are shown in Table 8. Regarding the regression-based evaluation, the system with an ensemble learning model (Run 1) has obtained 0.325 in RMSE, achieving the best results of all submissions. With regard to the ranking-based evaluation, which is mainly based on early detection evaluation metrics, our system with the ensemble learning model has reached the 10th position with a value of 0.333 on p@30. Our carbon emission for this task in both runs has a mean duration of 19.49339786 with a deviation of 19.88332675 and a mean emission of 5.52E-08 with a deviation of 5.44E-07.

Figure 5 shows a comparative graph between our best run (Run1) and the actual values. We can see that both lines have the same trends and that there are many cases with a value of 1 and our model predicts them with a value higher than 0.8, so, as an improvement measure, we

Table 8
Results of UMUTeam for Task 2.b in regression-based and ranking-based evaluation. For each run, RMSE and Pearson coefficient are reported for decision-based evaluation and p@5, p@10, p@20 and p@30 for ranking-based evaluation.

	RMSE	Pearson coefficient	p@5	p@10	p@20	p@30
Run 0	0.333	0.484	0.400	0.200	0.350	0.300
Run 1	0.325	0.522	0.400	0.500	0.350	0.333

can set a threshold in our model to assign a value of 1 when it is exceeded.

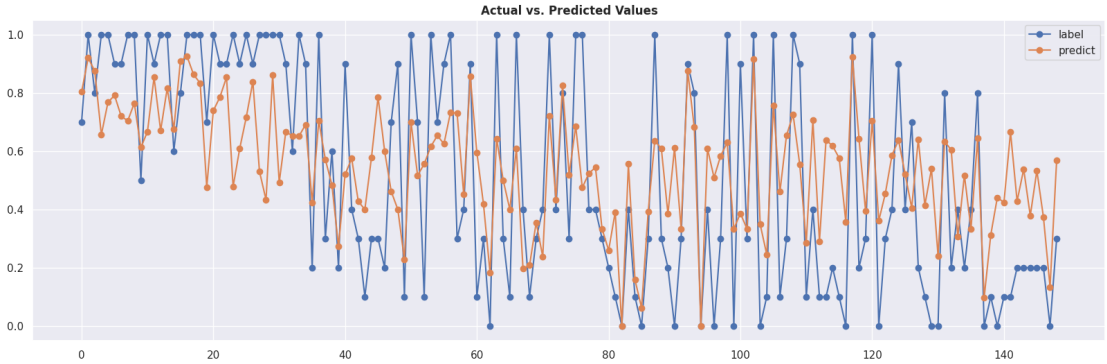

Figure 5: Comparison chart between actual and predicted values for Task 2.b.

4. Conclusion

This paper describes the participation of the UMUTeam in all the subtasks of Task 1 and Tasks 2.a and 2.b of the MentalRiskEs@IberLEF 2023 edition. This task aims at the early detection of eating disorders, depression, and undefined disorders that are related to eating disorders and depression through Spanish comments from Telegram users. For this purpose, the organizers are faced with two different perspectives: as a binary classification and as a simple regression problem. However, in the task of early detection of depression (Task 2), in addition to considering it as a binary classification and simple regression problem, they have also defined it as a multi-class classification and multi-output regression problem. For both Task 1 and Task 2, a fine-tuning approach of pre-trained monolingual and multilingual models has been proposed for the binary classification problem. An ensemble learning approach, concatenating the pooler output of the pre-trained model with the last hidden state of a pre-trained model for emotion identification, was also tested. As for the regression problem, instead of building a model from scratch, the obtained classification models were used, adding a softmax transformation to the output to obtain the probability of suffering. It should be noted that the conservative strategy was used to deal with the early detection problem, which consists in processing all the user's comments to make a decision.

In Task 1.a (binary classification), we have obtained good results, reaching the second position with a macro F1 of 0.918 in the decision-based evaluation and the 13th in the latency-based raking. As for the simple regression problem of Task 1 (Task 1.b), our system with the ensemble learning model has obtained the ninth-best result of all those presented with an RMSE of 0.255 in the regression-based evaluation and 14th in the ranking-based evaluation. As for Task 2, we obtained a better result in the binary classification problem with a macro F1 score of 0.737 in the decision-based evaluation and 31st in the latency-based evaluation. In the simple regression problem of Task 2 (Task 2.b), our system with ensemble learning obtained the fifth-best result with an RMSE of 0.325 in the regression-based evaluation and 10th in the ranking-based evaluation.

From the results obtained, it can be seen that it obtains very good results in predicting whether a user suffers from eating disorders or depression, but it does not achieve good results in early detection, because in this case we used the conservative strategy of processing all the user's comments to make a decision.

In future work, we plan to analyze other early detection techniques to improve the performance of Task 1 and Task 2. In the error analysis of Task 1.b and 2.b (see Section 3.1 and 3.2) it was found that our system usually assigns a value greater than 0.8 to cases that have a value of 1. Therefore, we can set a threshold (e.g., a value greater than 0.8) for the system to identify it as a value of 1, and this would improve the performance of the system. Moreover, we consider that it is relevant to consider the relationship between depression signs and hate-speech [8], usage of humour [9], and demographic and psychographic traits of the authors of the messages [10].

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

[1] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: Early risk prediction on the internet, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 343–361.

[2] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del-Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalriskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[3] J. M. Pérez, J. C. Giudici, F. Luque, pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks, 2021. arXiv:2106.09462.

[4] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 1–10.

[5] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, Albeto and distilbeto: Lightweight spanish language models, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, "Marseille, France", 2022, pp. 4291–4298.

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

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[8] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.

[9] J. A. García-Díaz, R. Valencia-García, Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers, Complex & Intelligent Systems 8 (2022) 1723–1736.

[10] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.