Improving Transformer by Instance Packaging for Mental Illnesses Identification

María de-Jesús-García-Santiago[1,*,†], Fernando Sánchez-Vega[1,2,†] and Adrián Pastor López-Monroy[1,†]

[1] *Mathematics Research Center, Jalisco S/N Valenciana, 36023 Guanajuato, GTO México*

[2] *Consejo Nacional de Ciencia y Tecnología (CONACYT), Av. de los Insurgentes Sur 1582, Benito Juárez, 03940, CDMX, México.*

Abstract

This work is an overview of the *CIMAT-NLP* submission at the shared tasks of MentalRiskES at IberLEF 2023, which consisted of three different tasks related to the detection of mental illness on social media text in Spanish: detection of eating disorders, depression, and an unknown disorder. In this work, we proposed using models based on RoBERTuito using two ideas for their training and evaluation. The main idea of our work is to make packages of messages text to create new instances with more information, as the messages provided in the training data are small in length words. Our proposed models were first on precision metrics in Task 1b and fifth in RSME on same subtask. In depression Task 3a our model placen fifth in Macro-F1 and second on ERDE30.

Keywords

mental disorder detection, depression, eating disorders, natural language processing

1. Introduction

A mental illness (or mental disorder) is a disease that involves different alterations in thinking, emotions, or behavior, the mental disorders are usually related to stressful events or problems functioning in social, work, or family activities [1].

One in eight persons suffered from a mental disease in 2019, where the most common disorders were depression and anxiety, the number of persons affected by these disorders incremented during the COVID-19 pandemic [2]. Even though mental diseases are common in the population, persons with these diseases have been stigmatized and discriminated against by society. Notwithstanding the fact that, mental illness could be prevented and diagnosed on time, many persons can not access effective treatments and diagnose for their disease; this could be because lack of information about mental illness, inefficient health systems, economic problems or help from their near circle [3].

Automatic detection of mental illness is a complex problem because of different factors; the complexity of the mental illness to detect, the data available for the training of the models, the

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

† These authors contributed equally.

maria.garcia@cimat.mx (M. de-Jesús-García-Santiago); fernando.sanchez@cimat.mx (F. Sánchez-Vega); pastor.lopez@cimat.mx (A. P. López-Monroy)

tools developed for this task and other factors.

The problem of automatic detection of mental disorders worsens in Spanish speaking countries. The predomination of the tools, programs, and models for the task is for the English language. In response to the absence of automatic systems to detect mental disorders in Spanish, MentalRiskES is a novel task on early risk identification of mental disorders in Spanish comments. With this novel task, the teams intended to do an online detecting duty, where the point was to detect a potential risk as early as possible, by simulating that the systems received messages like in a normal conversation.

Our proposed work has two main ideas: (1) Packaging of text for training and evaluation and (2) data augmentation. We used models based on transformers pre-trained in Spanish social media corpus, which is RoBERTuito [4, 5, 6]. We used the packaging idea for the training and evaluation of the data, as packages give more context to the model and the performance increments. In the second proposal, we still use the idea (1), adding the second idea (2) to augment the training data, as the package makes fewer instances for training. The augmentation of data is per user; as we make more instances of packages of each of them, the samples of messages in their records are used to make the packages and then used as training instances.

The rest of this paper is organized as follow: In Section 2 we introduce history about the detection of mental illness on social media. Section 3 describes the competition and the distribution of the datasets. In Section 4 we talk about our approaches for the different tasks and in Section 5 we decide the models to use for each task with their results on the competition.In Section 6, we aboard the ethical issues of the competition and the environmental impact. Finally, Section 7 contains the conclusion of this work.

2. Related work

In mental illness detection, one of the essential workshops is the CLEF eRisk: Early risk prediction on the Internet. The main objective is to develop automatic systems that alert depending on the task. Because of the importance of this workshop, most of the world had been devoted to the English language, but now MentalRiskEs tackles Spanish for the first time.

The majority of the proposals in eRisk since 2017 are Bag of Word-based or Neural Network based. In more recent editions, such as in 2020, the transformers started being used; one example in [7] in which the authors used XLM-RoBERTa [8] for the detection of depression and self-harm and obtained the first place for early detection of self-harm and in the top five for detection of depression.

Other works using transformers models are had been at the top of the Shared task on Detecting Signs of Depression from Social Media Text (LT-EDI) contest in 2022. The LT-EDI contest focuses on depression detection without the early detection factor, the works of [9, 10, 11, 12] that achieve the better performance at multiclass classification, used diverse transformers based of models such as as BERT [13], RoBERTa [14], DistilBERT [15], ELECTRA[16], DeBERTa[17], ALBERT[18], T5[19]. These works use different techniques in order to improve the performance of their proposed models. The techniques vary from fine-tuning in specific domine, ensembles, Valence Aware Dictionary for sEntiment Reasoning (VADER), which is used to make VAD scores and mask words in the training and augmentation data techniques (Back translation as one example of this).

In MentalRiskEs, the messages have different lengths; the majority of these messages could be considered short messages. The problem is that these messages, as instances, could not be enough to make decisions about the user; context is essential for model languages as they help to make a decision based on data. The messages may not contain relevant information and could mislead the users' emotional state. Because of the previous ideas, we hypothesize that the union of several messages to be treated as one would give the necessary context for a better decision.

An approach to the problem of short messages is in the work of [20]. This work references transformer models in Spanish tasks; in their work, they have a novel approach for the Author Profiling problem by predicting the labels at the tweet level instead of the user level. The tweets could be only one tweet or three packages for the predictions. We use this idea of packages of messages in our models proposed in Section 4. We think the packages provide more context to the transformer as the individual messages are usually short; this lead to a better performance of the models. One of the transformer's problems is the need for big datasets for their training, as mentioned in [21], where they use a packaging strategy to reduce the training time. Because the original number of instances was small for all the datasets, we propose an augmentation data technique using the idea of packages, which uses random messages from the history of each user.

3. Dataset and taks

The datasets of MentalRiskES are data consisting of messages from telegram users. The eating disorder and depression sub-datasets have the same number of users: 10 for trial, 175 for training, and 150 for testing. The third task only has 150 users for testing, as the organizers intend that the teams use their subsystems developed in the previous tasks.

As the primary purpose of the competition is to predict the risk of mental illness the earliest possible, they have a server where each team communicates to get the test data. Because they wanted to simulate a conversation in real-time, we have one message per-round and predict the possibility of each user having a mental illness. In each round, we got a list of users with the id_messages, names (for example, "subject1"), messages, and dates of the messages.

3.1. Task 1: Eating disorders detection

The first task intention was to detect if users suffer from bulimia or anorexia. This problem is sub-divided into two sub-tasks,

a The first subtask was a binary classification: only predicting one of the two labels ("suffer" or "control").
b The second subtask was simple regression: the probability of suffering the mental illness.

For the distribution of suffer and control users see Table 1, the same datasets were used for both subtasks.

	Training	Trial	Test
Total users	175	10	150
Suffer label	74	5	64
Control label	101	5	86

Table 1
For the task we have three datasets, the training and trial data was used for the development of the systems. Training and Test data have more control users than suffer users

3.2. Task 2: Depression detection

This problem is sub-divided into four sub-tasks,

a Binary classification: only predicting one of the two labels ("suffer" or "control").
b Simple regression: the probability of suffering the mental illness.
c Multiclass classification, only predicting one of the four classes (“suffer+against”, “suffer+in favor”, “suffer+other”, “control”).
d Multi-output regression: as simple regression but the probability corresponding to each class.

Similar to Task 1, the subtasks uses the same dataset for training and test.In Table 2 we only show the distribution respect if a user has or not the mental illness, because we did not participate in subtask 2c and 2d.

	Training	Trial	Test
Total users	175	10	150
Suffer label	94	6	68
Control label	81	4	82

Table 2
The number of user in each dataset remains as the imbalance noted (control users are more than suffer users).

3.3. Task 3: Unknown disorder detection

The last task was to detect an unknowing disorder (anxiety). Because we did not have a training corpus and previous knowledge about the disorder to be detected, we used the training corpus from Task 1 and Task 2 to make a model that detects general mental illness. This task has identical subtasks to Task 1. In this task, the number of users with the "suffer" label is 93, and the "control" label is 57.

4. Method

Because the dataset is a Spanish corpus, we also use transformers pre-trained in this language. RoBERTuito is a pre-trained model for content in Spanish [4, 5, 6]. The training used for

RoBERTuito was the guidelines of RoBERTa on 500 million tweets with the Whole Word Masking technique [22].

As RoBERTuito is based on RoBERTa, the range of text length input is a maximum of 768 tokens. The training dataset from MentalRiskEs has messages with small lengths. If we join all the messages from each user, most users' complete history has a mean of 500 tokens. Almost all users' complete history can be given to RoBERTuito without truncation. The main problem with this approach is that all the records have different lengths. The difference between lengths could lead to the model specializing in learning a fixed number of the tokens (the more repeated lengths), and the user's objective could not be the best approach, as we have small information in each round.

In our approach, we decided to make packages from joined messages for the training and validation dataset as in [20] and then use a voting scheme to predict the label for each user. In the second subsection, we make augmentation data for more instances to train the transformers and improve their performance.

4.1. Package Approach

In the training and validation data, a fixed size of messages is joined into packages because most messages in the competition corpus have a small length. When we merge the messages into packages, we add more information to the instance. Consequently, the model would be better for predicting the label for each package in the validation data. The order of the joining is date order, from the earliest to the latest, as we hypothesize that the temporal information is helpful as they simulate the user's mood during their history of messages.

For the prediction of the users of validation data, even though the predictions are individual, the user's decision is made by a voting scheme, which means that if the model predicted more often "suffer" labels, then the user will be labelled as "suffer," or in another case is labelled as "control."

In Fig.1, we have illustrated the packaging strategy described before for one user. First, the user has N instances, and then taking k subsequent messages, we join them in one package, so the number of packages will be N/k.

Instead of predicting all the messages, the model predicts the N/k instances created and finally makes one prediction for the user using the voting scheme and the new instances, the process is represented in Fig. 2.

For the experiment, we tried different learning rates and train batch sizes, where the evaluation dataset is the new validation dataset mentioned before. The size of the packages tried were 3, 5, and 10 for the training dataset following the same ideas in [20]. For the new validation dataset, the sizes were 2, 3,4, 5, and 10 as in the competition, and we do not know the lengths of the messages for a round. In Table 3, we show the new distribution of instances for the training and validation data in both tasks using the packages strategy.

One problem that surges with the packaging strategy is the reduction of instances for training. The number of instances is essential because the transformer models usually perform better when they have big datasets instead of small ones. To solve this problem, we developed a technique for augmentation data.

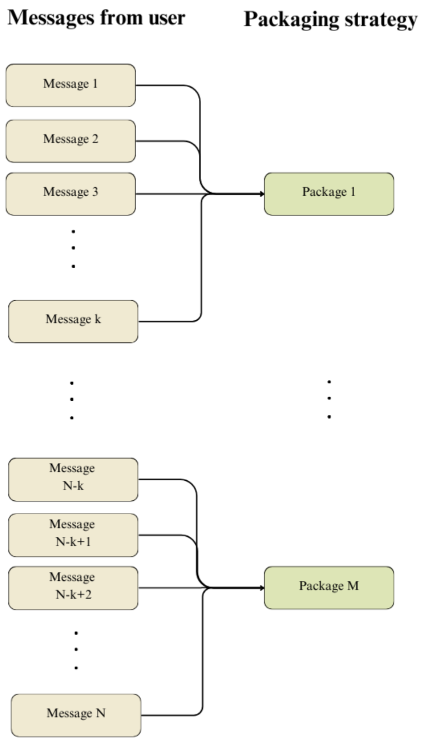

Figure 1: The packaging strategy is only made per user and not taking all the messages from all the users to make the packages, and this is because we want to ensure that the new instances have the correct label and to keep a consistent writing style.

	Task 1		Task 2	
	Train	Val	Train	Val
Original	4748	1572	5330	1542
Packages 2	-	790	-	778
Packages 3	1630	538	1822	527
Packages 4	-	490	-	395
Packages 5	993	322	1118	321
Packages 10	526	166	595	168

Table 3
Note that in this table, we are talking about individual messages instead of the number of users, the "original" dataset refers to the datasets without the packing strategy, and the following rows have the resultant instances after the strategy.

4.2. Augmentation data technique

For the augmentation data technique, we still used the process of making packages as in the previous subsection. The new part of this process is to make the packages with random messages from the same users to make more instances. As in Subsection 4.1, the packages are made

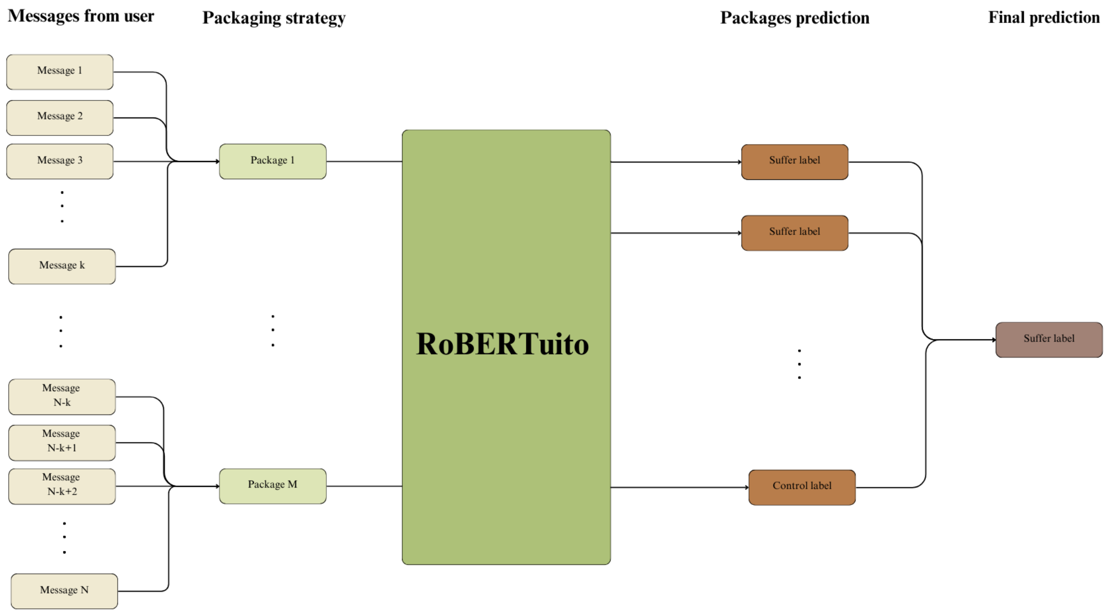

Figure 2: The packaging strategy is only made per user in the validation set. Each package is passed on to the encoder of RoBERTuito, and for the inference part, we obtain the prediction of the package. Finally, our final prediction is made using a voting scheme.

from messages from the same user intending to preserve the user's writing style, and the new instances are correctly labelled.

First, we make the packages in the process mentioned in the Subsection 4.1. Then, the packages for the augmentation data have messages selected randomly without the date as an essential factor for the joining to create new instances that are not duplicates and could contain different contexts. The final number of instances per user would be $N/k + (N/k) * r$, where N is the total of messages of the user, k the size of the package and r the number of times we repeat the process of making "random" packages.

Our main reason for making augmentation data is to give our model instances which contain messages that are from different moments in the life of the user and help to understand the patterns of their mood. Another benefit of this technique is that this solves the problem of working with small data, as the new datasets are more extensive and improve the model's performance. The same values used in Subsection 4.1 were used to make the augmentation datasets for training; the number of instances for these datasets triplicates the original number of packages shown in the Subsection 4.1; the validation datasets remained the same. Table 4 shows the distribution of the augmented datasets.

5. Results

This Section contains two subsections. In the Subsection 5.1, we talk about the results obtained in the validation datasets created. These results helped us to choose the best hyperparameters

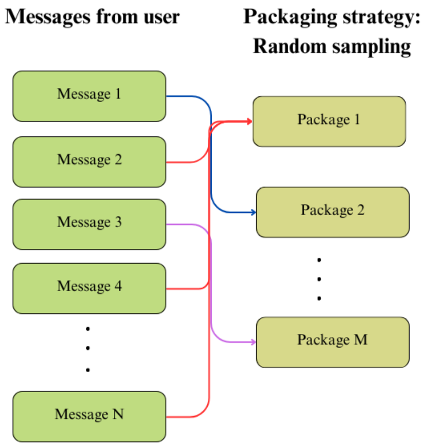

Figure 3: The packaging strategy is only made per user and not taking all the messages from all users to make the packages; as we see in the image, the packages are made of randomly selected messages, and the thing they have in common is the number of messages per package.

	Task 1		Task 2	
	Train	Val	Train	Val
Original	4748	1572	5330	1542
Packages 3	4889	538	5465	527
Packages 5	2978	322	3353	321
Packages 10	1577	166	1984	168

Table 4
The training dataset made with packages triplicate their size in comparison to the ones showed in Table 3. Eventhough they share one part of their data, as both type of dataset have instances created by the process described in Subsection 4.1

for the models of each task. The Subsection 5.2 shows the results obtained from our proposed models on the competition.

5.1. Models validation and hyperparameters selection

The F1-score binary metric for the classification of the users was used to select the best models for the competition.

Since the size of the trial data is only ten examples, the train and trial data merged into one and split into validation data with 20% of the data. For the experiments, we used the uncased version of RoBERTuito [4].

5.1.1. Task 1

For the detection of eating disorders we made experiments using different hyperparameters: learning rate and batch training size for the RoBERTuito transformer. First, we are going to

		Packing size for validation datasets					
Pack size for train	**Batch size train**	**Pack of 1(Original)**	**Pack of** 2	**Pack of** 3	**Pack of** 4	**Pack of** 5	**Pack of** 10
3 packages	32	0.9142	0.9142	0.9142	0.9142	-	-
	128	0.9696	1	0.9696	0.94	-	-
5 packages	**16**	**0.9375**	**1**	-	**1**	**1**	-
	128	0.9142	0.9411	-	0.9696	0.9696	-
	32	0.9142	0.9411	-	0.9696	0.9696	-
10 packages	32	0.8387	0.9696	-	0.9411	-	0,9696
	32	0.8888	0.9142	-	0.9411	-	0.909
	16	0.9696	0.9696	-	0.9393	-	1

Table 5
The column of Pack size for train refers to the size of packages made in the Packaging strategy and later used for the training. The follow columns refers to the validation datasets. "-" is because the model was not tested in that validation dataset. The model chosen for the competition is marked in bold in the column of batch size.

present the best models trained with our approach, and ranking respect F1-score metric. Later, we present the best models trained with our approach adding the augmentation technique and raking respect F1-score on validation dataset.

In Table 5, all the models are RoBERTuito fine tuned in the training datasets mentioned in Subsection 4.1. The first row present a model which is the only one that was trained with learning rate $1e^{-5}$, the others model are trained with learning rate $5e^{-5}$, and the hyper-parameter that changes is the batch size for training.

In Table 6, the results of the augmentation data technique are shown. The models' performance with this augmentation data is better than the models' performance only trained with the package strategy.

The models chosen were,

- **RoBERTuito-T5-B16**: with learning rate $5e^{-5}$ and batch size 16 trained with packages of 5.
- **RoBERTuito-T5-B16-DA**: with learning rate $1e^{-5}$ and batch size 16, this model was trained with augmented data and in packages of 5.

Because this were the models that had better F1-score in the majority of their validation datasets, we decide to use packages of 3 for the test dataset in the rounds, as the performance increase with the use of more information.

For the Task 1b the logits of the output prediction from the models were used as the prediction for the subtask.

5.1.2. Task 2

The results of the depression detection task are shown in Tables 7 and 8, as in Task 1 the F1-score binary was intended to be maximize.

In the case of depression, most of the learning rate used was $1e^{-6}$, contrary to the eating disorder task. This learning rate could be because the model needed more steps to learn

Pack size for train	lr	Packing size for validation datasets Batch size train	Pack of 1(Original)	Pack of 2	Pack of 3	Pack of 4	Pack of 5	Pack of 10
3 packages	$5e^{-5}$	16	0.89	0.967	1	0.9411	-	-
	$1e^{-5}$	128	0.9696	1	1	0.9696	-	-
	$1e^{-5}$	64	0.9375	1	1	0.9696	-	-
5 packages	$1e^{-5}$	**16**	0.8823	0.9696	-	0.9696	**1**	-
	$5e^{-5}$	32	0.9696	0.9375	-	1	0.9696	-
	$5e^{-5}$	16	1	1	-	1	1	-
10 packages	$1e^{-5}$	32	0.9142	1	-	0.9375	-	1
	$1e^{-5}$	16	0.9142	1	-	0.9696	-	1
	$5e^{-5}$	128	0.9677	0.9677	-	0.9677	-	1

Table 6
The first column refers to the training datasest created as described in Section 4.2, the other columns refers to the validations datasets made with the process in Section 4.1. We note that the results with augmentation data have better performance that results showed in Table 5. The model chosen for the competition is marked in bold in the column of batch size.

Pack size for train	lr	Packing size for validation datasets Batch size train	Pack of 1(Original)	Pack of 2	Pack of 3	Pack of 4	Pack of 5	Pack of 10
3 packages	$1e^{-6}$	128	0.7692	0.8	0.8163	0.8085	-	-
	$5e^{-5}$	64	0.7142	0.8085	0.7999	0.7999	-	-
	$1e^{-6}$	128	0.7692	0.8	0.8163	0.8085	-	-
5 packages	$1e^{-6}$	**32**	0.8181	0.8780	-	0.8648	0.8292	-
	$1e^{-6}$	128	0.8444	0.85	-	0.8292	0.8947	-
	$1e^{-6}$	64	0.8181	0.8205	-	0.8947	0.8947	-
10 packages	$1e^{-6}$	32	0.8	0.888	-	0.9	-	0.7692
	$1e^{-6}$	32	0.8163	0.909	-	0.871	-	0.75
	$1e^{-6}$	16	0.8181	0.8717	-	0.8333	-	0.7

Table 7
The structure of this table is the same to Table 6. The model chosen for the competition is marked in bold in the column of batch size.

adequately the patterns of persons with depression. In both techniques, only one model had different learning rates but worse performance than the others.

The models chosen were,

- **RoBERTuito-T5-B128-DA**: with learning rate $1e^{-6}$ and batch size 128, this model was trained with augmented data and in packages of 5.
- **RoBERTuito-T5-B32**: with learning rate $1e^{-6}$ and batch size 32 trained with packages of 5.

The reason for choosing the previous models is because the ERDE5 and other metrics that measure early detection, as the other models, have better performance with packages of information with more messages, and model the model performs better with packages of 2. Consequently,

Pack size for train	lr	Packing size for validation datasets						
		Batch size train	**Pack of 1(Original)**	**Pack of** 2	**Pack of** 3	**Pack of** 4	**Pack of** 5	**Pack of** 10
3 packages	$1e^{-6}$	128	0.6874	0.7222	0.7179	0.8947	-	-
	$1e^{-6}$	64	0.6451	0.7222	0.6666	0.8648	-	-
	$1e^{-6}$	32	0.625	0.6857	0.666	0.7272	-	-
5 packages	$1e^{-5}$	32	0.7272	0.6857	-	0.8	0.7894	-
	$1e^{-6}$	**128**	0.7567	0.8333	-	0.8333	0.8180	-
	$1e^{-6}$	64	0.6874	0.7272	-	0.8108	0.8333	-
10 packages	$1e^{-6}$	128	0.8095	0.8333	-	0.8648	-	0.7567
	$1e^{-6}$	64	0.7692	0.7843	-	0.9090	-	0.8095
	$1e^{-6}$	64	0.5625	0.7692	-	0.7906	-	0.80

Table 8
The structure of this table is the same to Table 6. The augmentation data technique was used for the training phase. The model chosen for the competition is marked in bold in the column of batch size.

we decide to use packages of 3 for the test dataset in the rounds, as the performance increases with more information.

For Task 2b, the logits of the output prediction from the models were used as the prediction for the subtask.

5.1.3. Task 3

Because Task 3 uses the subsystem of Task 1 and Task 2, the run A is an ensemble of two models:

- **RoBERTuito-T3-B16**: with learning rate $5e^{-5}$ and batch size 16, this model was trained with augmented data from eating disorders and in packages of 3.
- **RoBERTuito-T5-B32** with learning rate $1e^{-6}$ and batch size 32, this model was trained with data from depression in packages of 5.

5.2. Results in the competition

In this subsection, we show our team's results in the competition in the different subtasks. The metrics we use for comparison are F1-macro and ERDE30 for binary subtasks and RMSE, p@a, p@10, and p@20 for regression subtasks.

5.2.1. Task 1

In Task 1 we focused to increment the Macro-F1 for the eating disorders detection.

The results from model RoBERTuito-T5-B16 are reported; the model RoBERTuito-T5-B16-DA did not make it top. Table 9 compares with other teams from the official overview [23]. Team CIMAT-NLP-GTO and UMUTeam are included because they have the best performance at Macro-F1 score. Three BaseLine models from the committee organization are included for comparison.

For Task 1b, we provided the probability of the user suffering an eating disorder, and then the Root Mean Squared Error (RMSE) was calculated respecting the truth golden. The rest of the metrics are the precision of the prediction at 5, 10 and 20 messages, respectively.

	Task 1a		Task 1b			
	Macro-F1	**ERDE30**	**RMSE**	**p@5**	**p@10**	**p@20**
RoBERTuito-T5-B16	0.820	0.088	0.229	**1.000**	**0.900**	**0.900**
1st place	**0.966**	**0.018**	0.274	**1.000**	**0.900**	**0.900**
8th place	0.847	0.065	0.192	0.600	0.500	0.550
2nd place	0.918	0.113	0.257	0.600	0.700	0.650
BaseLine-Deberta	0.813	0.083	0.231	0.800	**0.900**	0.850
BaseLine-Roberta Large	0.813	0.099	0.196	0.800	0.800	**0.900**
BaseLine-Roberta Base	0.694	0.132	**0.178**	**1.000**	0.800	0.850

Table 9
Comparison between our results and another teams in both subtasks.RoBERTuito-T5-B16 was ranked 10 by Macro-F1 metric, and ranked 9 by the ERDE30 metric. In Task 1b our model placed fifth and in the other metrics placed first.

Even though our model was not the best in the Macro-F1 metric, it has competitive metrics for Task 1b, surpassing the models proposed as Baseline in almost all the metrics as they obtained first place on precision metrics. In the RMSE our model placed fifth, lower RMSE means better performance. Considering the hard baselines our team placed second and fourth in Task 1b.

5.2.2. Task 2

In the depression detection task, we focused too on maximizing the Macro-F1.

The results from model RoBERTuito-T5-B128-DA were in the top 20 of the results for Task 2, but the model RoBERTuito-T5-B32 was not at this top. Table 10 is shown the comparison with other teams from the official overview [23]. Team UMUTeam and UNSL are included because they have the best performance at the Macro-F1 score. Three BaseLine models from the committee organization are included for comparison. SINAI-SELA is added because the ERDE30 metric, CIMAT-NLP-GTO, was added by their performance in Task 2b.

As we can see, the Macro-F1 score maximum ranges are lower for eating disorders results. This is because depression detection usually is more difficult for a variety of themes that users with depression talk about.

In this case, the model using data augmentation was better that the model without it, we theorized that this technique is better with depression detection tasks, as they need more instances to classify correctly and we can note that our model had better performance in the ERDE30 than the best model in Macro-F1 or competitive results in this metrics in comparision to models with better Macro-F1.

5.2.3. Task 3

The results from model RoBERTuito-T3-B16 was in the top 15 of the results for Task 3. In Table 11 is shown the comparison with other teams from the official overview [23]. Team

	Task 2a		Task 2b			
	Macro-F1	**ERDE30**	**RMSE**	**p@5**	**p@10**	**p@20**
RoBERTuito-T5-B128-DA	0.645	0.187	0.335	0.600	0.600	0.550
1st place	**0.737**	0.358	0.333	0.400	0.200	0.350
2nd place	0.733	0.188	0.333	0.400	0.200	0.350
5th place	0.720	**0.140**	-	-	-	-
15th place	0.621	0.666	0.292	0.600	0.500	0.550
BaseLine-Deberta	0.813	0.083	0.231	**0.800**	0.600	0.550
BaseLine-Roberta Large	0.813	0.99	0.390	0.400	0.500	0.550
BaseLine-Roberta Base	0.694	0.132	**0.277**	0.600	**0.800**	**0.700**

Table 10
RoBERTuito-T5-B128-DA placed place twelve in the top 20 ranked by Macro-F1 metric in Task 2a, and placed seven in the top 20 ranked by RMSE metric in Task 2b.

CIMAT-NLP-GTO, NLP-UNED, and UPM are included because they have the best performance at Macro-F1 score and RSME. Three BaseLine models from committee organization are included for comparison.

	Task 3a		Task 3b			
	Macro-F1	**ERDE30**	**RMSE**	**p@5**	**p@10**	**p@20**
RoBERTuito-T3-B16	0.444	0.247	0.332	0.600	0.600	0.450
3rd place	0.650	0.285	0.482	0.200	0.300	0.350
11th place	0.402	0.231	0.435	0.600	0.500	0.600
1st place	**0.740**	0.188	0.348	**1.000**	**0.800**	**0.750**
BaseLine-Deberta	0.693	**0.165**	0.323	0.800	0.600	0.700
BaseLine-Roberta Large	0.630	0.179	0.374	0.200	0.100	0.350
BaseLine-Roberta Base	0.553	0.210	**0.308**	0.800	0.700	0.750

Table 11
RoBERTuito-T3-B16 placed place ten in the top 20 ranked by Macro-F1 metric in Task 3a and placed fourth in the top 20 ranked by RMSE metric in Task 3b.

The model proposed surpasses models with better Macro-F1 scores in task 3b, even though Macro-F1 is not the best; the model has competitive confidence in its predictions. Considering hard baselines, our team placed third and fifth in Task 3a, for the Task 3b, we placed on second place.

6. Ethical issues

The data used to develop systems for the automatic detection of mental illness is vital to emphasize the anonymity of the users' identities for whose text is recorded. Even though the identities are unknown, the labelling process is usually crowd-sourced; this process can not guarantee the correct labelling for the instances as it depends on the subjective judgment of the annotators. Furthermore, the data obtained did not consider the users' permission; these users could feel as if their privacy is being violated because it is crucial to use the data only to develop the systems.

6.1. Carbon emissions

Because of the critical climate change that is affecting the world is vital to make efforts to reduce the impacts of carbon emissions. The MentalRiskEs workshop implemented the tracking of the carbon emissions made by the models of the competition. They provided the code carbon package implemented on Python to track carbon emissions. The principal problem of this package with our models is that the functions did not track the emissions correctly, as they count all the GPU in the cluster instead of the only GPU used for the run of the models, and therefore calculate an incorrect amount of energy and carbon emissions produced. Another problem is the incompatibility with the version of Python greater than Python 3.8.

7. Conclusion

The MentalRiskES competition has been a challenging task concerning the early detection of mental health issues based on sequences of social media texts in spanish. Our models have balanced performance along all the metrics; their best performance is in the eating disorder dataset, which could be because of the nature of this dataset, as they tend to write more about their food-related issues. The depression dataset is more difficult by the complex writing of their authors and the need for more information on an excellent classification. We noted that for eating disorders, the model trained in the augmentation data did not have better performance than the model trained in the non-augmentation data; this could be because the model could be overfitting and making more false predictions of "suffering" class.

However, the depression detection models presented another pattern, the model trained with augmentation data was better for the task; we think this is because the model needs more information for depression users as they tend to write about an extended range of topics (not necessarily talk about their illness) like in eating disorder users.

For both tasks of linear regression, the models presented the best performances on the metrics about precision, which means that in the early stages of classification, the probability that the model assigned to each user is correct.

In future work, we plan to explore the size of the packages for the training and test datasets for eating disorder detection; for example, in each round, we could try to mix different packages for the test as we obtain more information or make different voting schemes. For the depression task, the augmentation data worked better than the other model. We plan to use the augmentation data we proposed to mix different package sizes to make new instances or clustering techniques to ensure that the messages joined have the same topic.

Acknowledgments

This research was funded by *Consejo Nacional de Humanidades Ciencia y Tecnología* (CONAHCyT) master's degree grant #1141296. The authors thank to CONACyT, CIMAT and *Instituto*

Nacional de Astrofísica, Óptica y Electrónica (INAOE) for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Plataforma de aprendizaje profundo para tecnologías del lenguaje*) and CIMAT Bajio Supercomputing Laboratory (#300832). Sanchez-Vega acknowledges CONACyT for its support through the Program "Investigadoras e Investigadores por México" by the project "Desarrollo de Inteligencia Artificial aplicada a la prevención de violencia y salud mental." (ID.11989, No. 1311).

References

[1] APS, What is mental illness?, 2022. URL: https://www.psychiatry.org/patients-families/what-is-mental-illness.

[2] W. H. Organization, Mental disorders, 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[3] National Collaborating Centre for Mental Health (UK), Common Mental Health Disorders: Identification and Pathways to Care, number 123 in NICE Clinical Guidelines, British Psychological Society (UK), Leicester (UK), 2011. URL: https://www.ncbi.nlm.nih.gov/books/NBK92265/, aCCESS TO HEALTHCARE.

[4] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

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

[6] M. García-Vega, M. Díaz-Galiano, M. García-Cumbreras, F. Del Arco, A. Montejo-Ráez, S. Jiménez-Zafra, E. Martínez Cámara, C. Aguilar, M. Cabezudo, L. Chiruzzo, et al., Overview of tass 2020: Introducing emotion detection, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, pp. 163–170.

[7] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Bert-based transformers for early detection of mental health illnesses, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 189–200. URL: https://doi.org/10.1007/978-3-030-85251-1_15. doi:10.1007/978-3-030-85251-1_15.

[8] 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, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[9] W.-Y. Wang, Y.-C. Tang, W.-W. Du, W.-C. Peng, NYCU_TWD@LT-EDI-ACL2022: Ensemble models with VADER and contrastive learning for detecting signs of depression from social media, in: Proceedings of the Second Workshop on Language Technology for

Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 136–139. URL: https://aclanthology.org/2022.ltedi-1.15. doi:10.18653/v1/2022.ltedi-1.15.

[10] R. Poświata, M. Perełkiewicz, OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 276–282. URL: https://aclanthology.org/2022.ltedi-1.40. doi:10.18653/v1/2022.ltedi-1.40.

[11] M. Janatdoust, F. Ehsani-Besheli, H. Zeinali, KADO@LT-EDI-ACL2022: BERT-based ensembles for detecting signs of depression from social media text, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 265–269. URL: https://aclanthology.org/2022.ltedi-1.38. doi:10.18653/v1/2022.ltedi-1.38.

[12] X. Lin, Y. Fu, Z. Yang, N. Lin, S. Jiang, BERT 4EVER@LT-EDI-ACL2022-detecting signs of depression from social media:detecting depression in social media using prompt-learning and word-emotion cluster, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 200–205. URL: https://aclanthology.org/2022.ltedi-1.27. doi:10.18653/v1/2022.ltedi-1.27.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[15] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[16] K. Clark, M. Luong, Q. V. Le, C. D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, CoRR abs/2003.10555 (2020). URL: https://arxiv.org/abs/2003.10555. arXiv:2003.10555.

[17] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced BERT with disentangled attention, CoRR abs/2006.03654 (2020). URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[18] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, CoRR abs/1909.11942 (2019). URL: http://arxiv.org/abs/1909.11942. arXiv:1909.11942.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, CoRR abs/1910.10683 (2019). URL: http://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[20] E. Villa-Cueva, I. González-Franco, F. Sanchez-Vega, A. P. López-Monroy, NLP-CIMAT at PoliticEs 2022: PolitiBETO, a Domain-Adapted Transformer for Multi-class Political Author Profiling, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings, CEUR-WS, 2022.

[21] M. Kosec, S. Fu, M. M. Krell, Packing: Towards 2x NLP BERT acceleration, CoRR

abs/2107.02027 (2021). URL: https://arxiv.org/abs/2107.02027. arXiv:2107.02027.
[22] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, CoRR abs/2111.09453 (2021). URL: https://arxiv.org/abs/2111.09453. arXiv:2111.09453.
[23] 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).