UNED-GELP at MentalRiskES 2024: Transformer-Based Encoders and Similarity Techniques for Early Risk Prediction of Mental Disorders

Jorge Fernandez-Hernandez[1], Hermenegildo Fabregat[1,3], Andres Duque[1,2], Lourdes Araujo[1,2] and Juan Martinez-Romo[1,2]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

[2]IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

[3]Avature Machine Learning, Spain

Abstract

This paper presents our participation in the MentalRiskES task of the IberLEF 2024 shared evaluation campaign, oriented to classification and early risk detection of mental disorders such as depression, anxiety and suicidal ideation. The UNED-GELP team has participated in tasks 1 and 3 of the competition, developing different systems for both tasks. In task 1, both our system based on approximate nearest neighbors and a different two-step system using Transformer-based encoding and K-Means classification achieve the third best results in multiclass classification among the participating systems. This last system also obtains the second best results considering early risk detection metrics for this task. Regarding task 3, our classification system trained on an external corpus and based on the combination of embeddings and linguistic features achieves the second best results for both the binary classification evaluation and the early risk detection metrics. An additional dictionary-based system has been developed for task 3.

Keywords

Early Risk Detection, Mental Disorders, Approximate Nearest Neighbors, Transformer-based encoders

1. Introduction

Mental disorders are one of the most important health problems worldwide, with an estimated 970 million people in the world currently living with a mental problem, including schizophrenia, depressive disorders, anxiety disorders, bipolar disorder, autism spectrum disorders, attention-deficit/hyperactivity disorder, conduct disorder, idiopathic developmental intellectual disability, eating disorders and other mental disorders, according to the World Health Organization [1]. Early detection of some of these conditions is crucial for a better understanding of the disease and for the development of accurate treatments. Beyond the usual therapies followed to diagnose and treat this type of disorders, nowadays the availability of large amounts of data coming from social networks and other similar sources allows for the development of automatic systems that enable early detection and much more agile clinical decision-making [2]. In this context, the MentalRiskES task [3] within the IberLEF 2024 shared evaluation campaign [4] aims to the development of automatic systems focused on the classification and early detection of various mental health issues such as depression, anxiety, and suicidal ideation, from textual messages obtained from the instant messaging platform Telegram, written in Spanish.

In this paper, we present our participation in the MentalRiskES task. We introduce various systems aimed at the classification and early detection of depression and anxiety, as well as the identification of users presenting suicidal ideation. These systems have in common the initial generation of embeddings for representing the analyzed messages. However, different techniques have been tested for the final

IberLEF 2024, September 2024, Valladolid, Spain

jfernandez@lsi.uned.es (J. Fernandez-Hernandez); gildo.fabregat@lsi.uned.es (H. Fabregat); aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)

0000-0001-9820-2150 (H. Fabregat); 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

classification of these embeddings, including models based on neural networks, approximate nearest neighbors, or K-means. Additionally, a purely unsupervised system based on dictionaries has been also employed for detecting suicidal ideation.

The rest of the paper is structured as follows: Section 2 is devoted to describe previous similar tasks and existing systems performing early detection of different kind of disorders through textual data analysis. The tasks and subtasks addressed, as well as the datasets involved in each of them, are explained in Section 3. Section 4 presents the different systems developed for each of the addressed tasks, while the results regarding these systems are gathered in Section 5. Finally, Section 6 offers some conclusions regarding this research as well as some possible lines of work to be followed in the future.

2. Related Work

Automatic early risk detection of different health problems such as anorexia, self-harm or pathological gambling has been explored in different editions of the CLEF eRisk workshop [5, 6, 7, 8]. Many different approaches can be found in these competitions, many of them based on Transformer architectures [9] for performing classification [10, 11, 12], as well as other deep learning models [13, 14, 15]. However, our proposal based on dataset relabeling from user-level annotations to message-level annotations through Approximate Nearest Neighbors obtained the best results in the 2022 competition [16] and the second best results in the 2023 competition [17]. Part of the research presented in this paper is an extension of that particular work.

Following the path set by the eRisk competitions, the MentalRiskES task was proposed in 2023 [18] within the IberLEF evaluation campaign [19]. In this task, Telegram is used as the data source instead of Reddit, which was used in the eRisk competitions. The subtasks considered in MentalRiskES 2023 included the detection of eating disorders, depression, and a third unknown disorder. Both the subtasks focused on detecting eating disorders and detecting the unknown disorder were tackled in terms of binary classification and simple regression, while the subtask that aimed to depression detection also incorporated multiclass classification and multi-output regression parts. In this competition, the best results were achieved by architectures based on Transformers [20, 21, 22]. In particular, in [20] classical approaches based on TF-IDF features and a Naïve Bayes classifier were compared against versions of RoBERTuito trained with different seeds. The work presented in [21] used the BETO model combined with a decision policy that took into account the prediction history of the model through the process of evaluating the user. Ensemble techniques involving different pre-trained Transformer models such as BETO, AlBETO, DistilBETO or XML-RoBERTa were employed in [22].

In the MentalRiskES 2023 competition, our system based on approximate nearest neighbors [23] achieved good results in the multiclass classification subtasks and acceptable results in the rest of the subtasks. Efficiency-related metrics were also considered in this competition, regarding energy consumption and CO2 emissions, as well as execution time. Our proposal ranked among the best according to all these additional metrics.

3. MentalRiskES Tasks

This section is dedicated to presenting the subtasks of the MentalRiskES competition in which we have participated, as well as information about the datasets used in those subtasks. The main evaluation metrics are also described at the end of the section. As mentioned before, the main objective of the MentalRiskES competition is the early risk detection of mental disorders using posts and comments from social media as a source of information. In particular, given a history of messages about a user in Telegram, the main aim is to detect, as soon as possible, whether the user suffers from a specific mental disorder. Therefore, a system will perform better the fewer messages it needs to predict that a user is at risk. The UNED-GELP team has developed systems for task 1 and task 3 of the competition.

3.1. Task 1: Disorder detection

In the first task of the MentalRiskES competition, systems are asked to determine whether a particular user suffers from anxiety, depression, or none of them, by analyzing the user's history of messages in Telegram. Hence, it is a multiclass classification task in which each user can be classified as belonging to one of the 3 possible classes: "depression", "anxiety", or "none". The organizers provided the participants with a trial dataset containing information about 20 users and a training dataset composed of 464 users. The dataset employed for testing purposes contained a total of 400 users. More information about the dataset can be found in [24].

3.2. Task 3: Suicidal ideation detection

In this task, systems are asked to determine whether each particular user presents symptoms of suicidal ideation, also by analyzing their Telegram history. We are then dealing with a binary classification problem, since the possible labels are 0 for "control" users (not presenting suicidal ideation symptoms) and 1 for "suffer" (positive users manifesting suicidal ideation symptoms). No trial or training data is provided for this task, hence the participating systems must either develop unsupervised systems or rely on their own or external data for training their systems. The test dataset is composed of 55 users.

3.3. Evaluation metrics

Two different subsets of evaluation metrics are proposed for this competition: performance metrics and efficiency metrics.

- **Performance metrics:** Classic metrics such as accuracy, recall, precision, and F1 score are employed for evaluating the systems' performance in terms of binary and multiclass classification. Additionally, a subset of metrics designed for analyzing the performance in terms of early prediction of the analyzed mental disorders is also proposed. These metrics include ERDE, speed, and latency, as well as latency-weighted-F1. More information about them can be found in [8].
- **Efficiency metrics:** The impact of the participating systems regarding use of resources and environmental issues is also measured through a subset of metrics. In particular, total RAM and percentage of CPU usage, Floating Point Operations Per Second (FLOPS), and total processing time are calculated. The carbon footprint is also evaluated by estimating the emission of CO2 (in kilograms), using the CodeCarbon tool [25].

4. Proposed Systems

This section is devoted to the description of the systems developed by the UNED-GELP team. As previously mentioned, our team has participated in task 1 and task 3 of the MentalRiskES competition, submitting three different runs to each task.

4.1. Task 1

As mentioned in Section 3, in task 1 of the MentalRiskES competition systems are asked to determine whether a particular user suffers from depression, anxiety, or none of them, by analyzing the user's history of messages in Telegram. Two different subsystems have been designed by the UNED-GELP team for addressing this task: the first system, employed in run 0, relies on the use of two models based on the transformer architecture, one for determining the existence of a mental disorder and a second one for classifying the disorder; the second system is based on the use of approximate nearest neighbour (ANN) techniques and has been utilised in run 1 and run 2 of the task.

4.1.1. Two-Step System

Two models are employed by this system in run 0 of task 1 for performing multiclass classification: the first model predicts whether a user is suffering of any mental disorder or not, this is, it determines whether the user should be classified as "positive". On the other hand, the second model only operates on positive users and determines whether they are suffering from either anxiety or depression. The two models operate in a cascading methodology. The idea behind this architecture is the following: it is very likely that the similarity between depressed and anxious users is higher (this is, messages are more similar to each other) than that between either of these two groups and negative users. This hypothesis led us to think that good results would be obtained by first building a classifier specialised in distinguishing between positive and negative users, which a priori have a lower similarity, and then a second classifier, which tries to be more refined since the similarity between users with depression and anxiety is a priori higher than in the previous case.

Each of the classifiers works under the same methodology: first a transformer-based model trained using sentence similarity techniques provides the embedding of the input text, which is then used by a K-Means model to give us the final prediction. In this model, two clusters are generated using the K-Means algorithm and then each cluster's label is assigned according to the majority class appearing in the cluster. The input text belonging to a particular user is obtained by concatenating all the user's messages for training. When performing prediction on the test set, for each new message the input text is created by concatenating all messages from a user that have been received by the system up to that point.

To train the first model, which discriminates between positive and negative users, all those users in the training dataset presenting either anxiety or depression were grouped under the same "positive" label. The rest of the users were labelled as "negative". A similar approach is employed with the second model, which predicts anxiety and depression. Negative users were excluded for the training process of this second model, and positive users were divided into two classes: those suffering from anxiety and those suffering from depression.

As previously stated, the models used to obtain the embeddings have been trained using sentence similarity techniques, through the *sentence-transformers* library[1]. The base transformer employed is BETO [26], a variant of BERT designed for processing Spanish texts. In order to train the models using sentence-similarity, the similarity of two sentences is computed using values between 0 and 1. In our case, each text, formed by the concatenation of all the messages of a user, is associated with two different sentences, one labelled with a similarity value of 1, and the other with a similarity value of 0. In particular, if a user is identified as positive, the input text is associated with the text "Esta persona tiene ansiedad o depresión" ("This person suffers from anxiety or depression") with a similarity value of 1, and also with the text "Esta persona no tiene ni ansiedad ni depresión" ("This person does not suffer from anxiety or depression") with a similarity value of 0. In the case of anxiety and depression, if the user has depression, the text is associated with "Esta persona tiene depresión" ("This person suffers from depression") with a similarity value of 1, and with "Esta persona tiene ansiedad" ("This person suffers from anxiety") with a similarity value of 0. If the user has anxiety, the similarity values are inverted.

80% of the provided training dataset was used for training and the remaining 20% for performance evaluation, using random selection, but maintaining the ratio between classes. Each of the models was trained with the default hyperparameters.

4.1.2. Approximate Nearest Neighbors

The second system used in task 1, and employed in both run 1 and run 2, is based on Approximate Nearest Neighbors (ANN) techniques, and is an extension of the work presented in the previous edition of the MentalRiskES task [23]. The main idea of this approach relies on performing a relabelling step on the training dataset. This way, we are able to shift from user-level labelling to message-level

[1] https://sbert.net/

labelling. First, we employ a Transformer-based encoder from the Universal Sentence Encoder family [27] for generating an embedding representing each message from a user. Then, all messages from a user are initially labelled with the same class assigned to that user (depression, anxiety or none). For the relabelling process, the approximate nearest neighbors library Annoy [28] is used. Through this technique, for each given message we can extract the J nearest messages in the search space, and analyze their labels. If the original message is positive, this is, labelled as "depression" or "anxiety", its label will be redefined as "none" only if a total of K messages out of those J originally retrieved messages belong to class "none". No relabelling is done for those messages originally labelled as "none", and messages labelled as one of the positive classes ("depression" and "anxiety") cannot be relabelled as the other positive class. The relabelling process is repeated until convergence is reached, this is, no new relabellings are performed during an iteration. Once the training dataset has been relabelled, the final classification step is done in a similar way to the relabelling process: two new J and K parameters are calculated in such a way that, given a new message from a user, its J nearest messages are retrieved from the training dataset, and the user class will be set to a positive class ("depression" or "anxiety") if at least K of those J nearest messages belong to the same positive class. Otherwise, the message (and hence the user up to that message) is classified as "none".

Regarding the two different runs of this system, run 1 employs the exact methodology previously described, while run 2 incorporates an improvement on the representation of the given messages. Once the relabelling process has been completed, a fine-tuning based on contrastive learning [29] is performed on the messages of the training dataset. This technique allows the system to improve the representation of the text messages by maximizing the distance between messages belonging to different classes and minimizing the distance between messages belonging to the same class. This is achieved through the use of a neural network that employs a contrastive loss function, such as triplet loss [30]. In our problem, for each message belonging to a particular class ("depression", "anxiety" or "none"), a triplet (a, p, n) is created were a is the embedding of the message, p the embedding of a message belonging to the same class, and n the embedding of a message belonging to any of the other two classes. The loss function defined for performing contrastive learning is $\mathcal{L} = max(d(a, p) - d(a, n) + margin, 0)$, where d is a function measuring the distance between the embeddings. Hence, the main objective of the network will be to minimize the distance between similar messages and maximize the distance between dissimilar messages. Parameter $margin$ determines the minimum distance that should exist between positive and negative instances, taking the original message a as a reference. This behaviour is illustrated in Figure 1, in which $margin$ has been renamed as α.

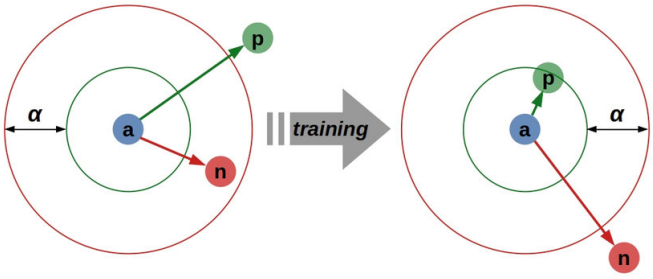

Figure 1: Result of the triplet loss function for performing contrastive learning: after training, the distance between similar (p) and dissimilar (n) instances with respect to a particular anchor instance (a) is maximized, by minimizing distance between a and p and maximizing distance between a and n.

The specific parameters of the contrastive learning technique implemented by this system are as follows:

- 20 triplets (a, p, n) are generated for each message a, by randomly selecting positive and negative instances.
- Batch size value is 32, while the learning rate is set to $1e^{-5}$.

- The number of epochs is 4.
- The number of steps per epoch is 128, this is, a maximum of 128*32 (batch size) instances are seen by the network in each epoch. Hence, a total of 128*32*4 = 16,384 instances are seen, i.e. not all triplets generated are finally seen by the network, and not all instances are seen the same number of times.
- The triplet loss margin is set to 0.15 (normalized values are used for distances and margin).

The rest of the pipeline regarding the relabelling process (previous to the contrastive learning process) and the inference step for performing the final classification is maintained similar to run 1. In both runs, 25% of the training dataset was reserved for validation and parameter optimization. The best values found for parameters (J, K) in the relabelling step and the final classification step for each of the runs were the following:

- Run 1: $(10, 5)$ for the relabelling process and $(10, 7)$ for the final classification step.
- Run 2: $(10, 5)$ for the relabelling process and $(11, 10)$ for the final classification step.

4.2. Task 3

Two different systems have been developed for addressing task 3, aimed at the identification of users with suicidal ideation. The first system is based on the Transformer architecture, enriched with extra features, and has been used in run 0 and run 1, while the second system is based on the use of an external dictionary and has been employed for run 2. As no training data is provided for this task, the training of the models, when needed, has been done on the SuicidAttempt corpus, which is focused on the identification of suicide attempts. The corpus is built from messages extracted from the instant message application Telegram, and consists of 146,733 messages, written in Spanish, from Telegram users. The corpus is annotated at user level, hence those users presenting an explicit mention of having commited a suicide attempt within their posts are annotated as positive, while those users not mentioning any explicit suicide attempt in their posts are annotated as negative. The corpus contains 150 positive users and 433 negative users for a total of 583 annotated users. More information about the corpus can be found in [31], and the most important characteristics are shown in Table 1.

Table 1
Statistics of the SuicidAttempt corpus.

Source	Language	# Messages	Positive Users	Negative Users
Telegram Groups	Spanish	146,733	150	433

4.2.1. Transformers with Extra Features

The architecture employed in runs 0 and 1 is based on the transformer architecture and incorporates additional features. In particular, it combines the embeddings obtained by feeding the Transformer-based model BETO [26] with the textual inputs, with the 200 terms with the highest average difference obtained from the TF-IDF algorithm. This is, for each term the difference between its TF-IDF value considering only positive and only negative users is computed, and those terms with highest difference values (those much more related to positive users than to negative users or vice versa) are selected as features. In the last step, a classification head is fed with this combination of embeddings and TF-IDF features for performing the final prediction. The architecture is illustrated in Figure 2.

Two different configurations have been designed for run 0 and run 1 of this particular task, both of them based on the aforementioned architecture. In run 0 the input of each particular user was built by concatenating all his messages. For run 1, on the other hand, the key message, i.e. the first message in which a positive user mentions that he/she has attempted suicide, is removed from the training set. This is done because the addressed task is focused on identifying ideation, while the corpus used for training is devoted to identifying suicide attempts.

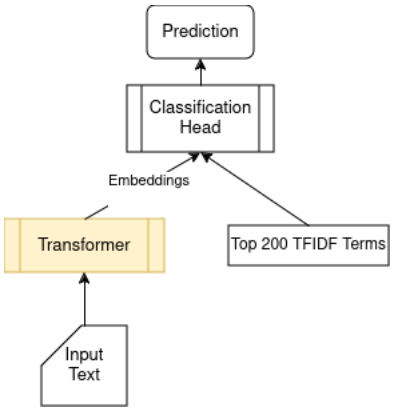

Figure 2: Architecture used in runs 0 and 1 of task 3: combination of Transformer-based embeddings and TF-IDF features.

A random subset containing 70% of the training data was employed for training the models, while the remaining 30% was left for evaluation purposes. However, the positive-negative ratio was maintained in the split. This methodology was found to be very sensitive to the initial value of the hyperparameters (especially the learning rate) during the training phase, hence a Bayesian optimization process was performed in order to look for the best combination of learning rate and batch size. The interval between $5e^{-6}$ and $9e^{-1}$ was considered for the learning rate, while the considered values for the batch size were 8 and 16.

4.2.2. Dictionary based system

Given that no training data was provided for this task, we were also interested in designing a system that did not require any training data. In particular, we developed a dictionary-based system. In this case, the dictionary was constructed from the lemmas more related to positive users in the SuicidAttempt corpus, according to the TF-IDF measure. The lemmas considered were: “suicidar” (“suicide”), “morir” (“die”), “cortar” (“cut”) and “pastillas” (“pills”).

For each new text belonging to a user, a pre-processing step including lowercasing, punctuation symbol removal and stopword removal is performed. Then, the lemmas of the remaining words are extracted. The system determines that a user is positive if more than 2 lemmas from the dictionary are found.

5. Results and Discussion

Main results obtained by the proposed systems, compared to other participants in the task, are shown in this Section. Regarding task 1, Tables 2 and 3 show the systems ranked by the Macro-F1 and ERDE30 metrics, respectively. As mentioned in Section 3, Macro-F1 is used for analyzing effectiveness in terms of multiclass classification, while ERDE30 is employed for analyzing early risk detection.

Table 2
Classification-based evaluation in Task 1. Metric ranking: Macro-F1. Shaded background highlights the baselines. Bold indicates the runs of the UNED-GELP team.

Rank	Team	Run	Accuracy	Macro-P	Macro-R	Macro-F1
1	ELiRF-UPV	2	0.890	0.875	0.880	0.874
2	ELiRF-UPV	1	0.850	0.853	0.845	0.840
3	BaseLine - Roberta Base	2	0.853	0.840	0.843	0.834
4	ELiRF-UPV	0	0.848	0.840	0.838	0.833
5	UnibucAI	0	0.828	0.824	0.808	0.808
6	UnibucAI	1	0.820	0.802	0.798	0.795
7	UnibucAI	2	0.820	0.808	0.793	0.793
8	**UNED-GELP**	**0**	**0.797**	**0.792**	**0.797**	**0.785**
9	**UNED-GELP**	**2**	**0.800**	**0.789**	**0.753**	**0.766**
10	Ixa-Med	1	0.790	0.796	0.747	0.749
11	**UNED-GELP**	**1**	**0.765**	**0.751**	**0.745**	**0.747**
12	Ixa-Med	2	0.790	0.790	0.733	0.736
13	Ixa-Med	0	0.762	0.763	0.725	0.723
14	BaseLine - Roberta Large	1	0.670	0.786	0.708	0.682
15	UMUTeam	2	0.690	0.701	0.683	0.675
16	UMUTeam	0	0.630	0.728	0.662	0.640
17	BUAP_ 01	1	0.620	0.692	0.662	0.632
18	BaseLine - mDeberta	0	0.710	0.748	0.645	0.623
19	UC3M-DAD	0	0.578	0.727	0.647	0.601
20	UC3M-DAD	1	0.578	0.727	0.647	0.601
21	UC3M-DAD	2	0.578	0.727	0.647	0.601
22	NLP UNED MRES	0	0.557	0.644	0.620	0.561
23	BUAP_ 01	0	0.427	0.650	0.557	0.411
24	BUAP_01	2	0.393	0.348	0.352	0.348
25	VerbaNex AI	1	0.527	0.598	0.372	0.303
26	VerbaNex AI	2	0.527	0.598	0.372	0.303
27	VerbaNex AI	0	0.512	0.551	0.353	0.271
28	UMUTeam	1	0.515	0.712	0.355	0.269
29	NLP UNED MRES	1	0.352	0.564	0.402	0.264
30	NLP UNED MRES	2	0.318	0.664	0.383	0.237
31	Huerta	0	0.470	0.240	0.318	0.231
32	Huerta	1	0.470	0.240	0.318	0.231
33	Huerta	2	0.470	0.240	0.318	0.231

Table 3
Latency-based evaluation in Task 1. Metric ranking: ERDE30. Shaded background highlights the baselines. Bold indicates the runs of the UNED-GELP team.

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

Our systems are able to obtain good results for both types of evaluation. In particular, our two-step based system (run 0) achieves the third best position among participating systems regarding Macro-F1 and the second best position regarding ERDE30, this is, the systems presents a good performance both in terms of classification and early detection. The two subsystems based on approximate nearest neighbors (either using contrastive learning fine-tuning or not), this is, runs 1 and 2, obtain similar results to run 0 in terms of Macro-F1, with run 2 slightly overcoming run 0 in terms of accuracy, however, their values of ERDE30 are higher than those achieved by run 0. Hence, their rank is quite lower in terms of early risk detection. This is probably due to the fact that run 1 and run 2 consider each message individually (since the training dataset has been relabelled at a message level), and hence analyze each new test message as it arrives. Run 0, on the other hand, analyzes the complete history of messages up to the latest message received and hence is able to deal with a higher amount of information about the user. This probably leads to an earlier detection of users at risk (smaller ERDE and latencyTP values and higher speed values).

It is also interesting to remark that the baseline implementing a RoBERTa base model is able to obtain the second best results in classification and the best results in early risk prediction, which indicates the difficulty in overcoming the performance of large models for this particular task.

Tables 4 and 5 illustrate the results obtained by the systems participating in task 3, suicidal ideation

detection.

Table 4
Classification-based evaluation for Task3. Metric ranking: Macro-F1. Shaded background highlights the baselines. Bold indicates the runs of the UNED-GELP team.

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

Table 5
Latency-based evaluation in Task 3. Metric ranking: ERDE30. Shaded background highlights the baselines. Bold indicates the runs of the UNED-GELP team.

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

In this task, our model trained on the SuicidAttempt corpus, combined with extra features extracted with the TF-IDF metric is able to obtain the second best results for both classification and early risk detection. In particular, run 0, which includes the key messages from the SuicidAttempt corpus for training the models, obtains the best results for the Macro-F1 and ERDE30 metrics, while removing these key messages (run 1) has a positive influence in terms of accuracy. The unsupervised system based on the use of a dictionary (run 2) offers lower results in both evaluations. The results of this system could be probably improved by increasing the range of terms considered in the dictionary. The dictionary-based model is very ineffective in terms of early detection since, as it searches for specific terms associated with suicide, it is unlikely that a positive prediction is done until an explicit message of suicidal ideation is found.

In this case, no system was able to beat the "all positives" baseline provided by the organizers for the early risk prediction metrics.

Finally, as mentioned in Section 3.3 different efficiency metrics have been employed for measuring

the environmental impact of the participating systems. Regarding CO2 emissions, our system generates 1.84×10^{-4} kg of CO2-equivalents for task 1, ranking 5th out of 10 participating teams. However, the amount of CO2 emissions for task 3 is higher, reaching 9.97×10^{-4} kg of CO2-equivalents. In this task our team ranks 4th out of 4 participating teams.

6. Conclusions and Future Work

This paper presents our participation in the MentalRiskES task within the IberLEF 2024 shared evaluation campaign. Different systems have been developed for each of the addressed tasks. In particular, in task 1 we propose a system based on the relabelling of the training dataset, from user-based to message-based labels, achieved through approximate nearest network techniques, also used for the final classification. An additional contrastive learning fine-tuning is also tested in this system. This system obtains the third best results under the multiclass classification metrics, compared to the rest of the participating systems. In this task, a two-step method that consider different models for detecting the existence of a disorder and for classifying the type of disorder obtains similar results in multiclass classification and is able to obtain the second best results according to the early risk detection metrics.

Regarding task 3, our best performing system relies on a Transformer-based encoder combined with the use of the 200 most informative terms according to TF-IDF, for performing the final classification. This system achieves the second best results regarding binary classification, and also in terms of early risk detection. Given the lack of training data for this task, our system has been trained on an external corpus devoted to detecting suicide attempts. A dictionary-based system has been also developed for this task, although its results are far from those obtained by the supervised system.

In general, the proposed Transformer-based encoders have proven useful for accurately representing the input texts. The contrastive learning process based on a neural network implementing a triplet loss function developed for run 2 of task 1 improves this representation by maximizing the distance between embeddings belonging to different classes. Once this input representation is generated and refined, we show that classic machine learning models such as approximate nearest neighbors, simple neural networks or even simpler models like K-Means are enough to obtain good results regarding the different metrics employed.

Future lines of work include the use of more sophisticated loss functions for performing the contrastive learning fine-tuning on the generated embeddings, as well as testing this techniques on additional problems and domains like the detection of other disorders such as anorexia, pathological gambling or self-harm.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, OBSER-MENH Project (MCIN/AEI/10.13039 and NextGenerationEU"/PRTR) under Grant TED2021-130398B-C21 and EDHER-MED Project under grant PID2022-136522OB-C21, as well as by the Universidad Nacional de Educación a Distancia (UNED) within project SICAMESP (2023-VICE-0029).

References

[1] W. H. Organization, et al., World mental health report: Transforming mental health for all (2022).

[2] C. Su, Z. Xu, J. Pathak, F. Wang, Deep learning in mental health outcome research: a scoping review, Translational Psychiatry 10 (2020) 116.

[3] 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 2024: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[4] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.
[5] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at CLEF 2020: Early risk prediction on the internet (extended overview), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2020 2696 (2020). URL: http://ceur-ws.org/Vol-2696/paper_253.pdf.
[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 2021 2936 (2021) 864–887. URL: http://ceur-ws.org/Vol-2936/paper-72.pdf.
[7] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet., Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy (2022).
[8] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet., Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece (2023).
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.
[10] A. Basile, M. Chinea-Rios, A.-S. Uban, T. Müller, L. Rössler, S. Yenikent, B. Chulví, P. Rosso, M. Franco-Salvador, Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet, Working Notes of CLEF (2021).
[11] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, Working Notes of CLEF (2021).
[12] A. M. Mármol-Romero, S. M. J. Zafra, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, A. Montejo-Ráez, SINAI at erisk@clef 2022: Approaching early detection of gambling and eating disorders with natural language processing, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 961–971. URL: https://ceur-ws.org/Vol-3180/paper-76.pdf.
[13] R. P. Lopes, Cedri at erisk 2021: A naive approach to early detection of psychological disorders in social media, in: CEUR Workshop Proceedings, CEUR Workshop Proceedings, 2021, pp. 981–991.
[14] H. Srivastava, L. N. S, S. S, T. Basu, Nlp-iiserb@erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 972–986. URL: https://ceur-ws.org/Vol-3180/paper-77.pdf.
[15] T. Dumitrascu, CLEF erisk 2022: Detecting early signs of pathological gambling using ML and DL models with dataset chunking, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 883–893. URL: https://ceur-ws.org/Vol-3180/paper-70.pdf.
[16] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, UNED-NLP at erisk 2022: Analyzing gambling disorders in social media using approximate nearest neighbors, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 894–904. URL: https://ceur-ws.org/Vol-3180/paper-71.pdf.
[17] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, NLP-UNED-2 at erisk 2023: Detecting pathological gambling in social media through dataset relabeling and neural networks, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Confer-

ence and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 672–683. URL: https://ceur-ws.org/Vol-3497/paper-056.pdf.

[18] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, L. A. U. López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish, Proces. del Leng. Natural 71 (2023) 329–350. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[19] M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Álvarez-Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3496.

[20] F. Echeverría-Barú, F. Sánchez-Vega, A. P. López-Monroy, CIMAT-NLP-GTO at mentalriskes 2023: Early detection of mental disorders in spanish messages using style based models and BERT models, in: M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Álvarez-Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3496/mentalriskes-paper16.pdf.

[21] H. Thompson, M. Errecalde, Early detection of depression and eating disorders in spanish: UNSL at mentalriskes 2023, in: M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Álvarez-Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3496/mentalriskes-paper4.pdf.

[22] R. Pan, J. A. García-Díaz, R. Valencia-García, Umuteam at mentalriskes2023@iberlef: Transformer and ensemble learning models for early detection of eating disorders and depression, in: M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Álvarez-Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3496/mentalriskes-paper9.pdf.

[23] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, NLP-UNED at mentalriskes 2023: Approximate nearest neighbors for identifying health disorders, in: M. Montes-y-Gómez, F. Rangel, S. M. J. Zafra, M. Casavantes, B. Altuna, M. Á. Álvarez-Carmona, G. Bel-Enguix, L. Chiruzzo, I. de la Iglesia, H. J. Escalante, M. Á. G. Cumbreras, J. A. García-Díaz, J. Á. G. Barba, R. L. Tamayo, S. Lima, P. Moral, F. M. P. del Arco, R. Valencia-García (Eds.), Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 26, 2023, volume 3496 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3496/mentalriskes-paper2.pdf.

[24] 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, MentalRiskES: A new corpus for early detection of mental disorders in Spanish, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics,

Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11204–11214. URL: https://aclanthology.org/2024.lrec-main.978.

[25] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, S. Luccioni, Codecarbon: Estimate and track carbon emissions from machine learning computing, https://github.com/mlco2/codecarbon, 2021.

[26] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[27] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal Sentence Encoder, CoRR abs/1803.11175 (2018). URL: http://arxiv.org/abs/1803.11175. arXiv:1803.11175.

[28] E. Bernhardsson, Annoy: Approximate Nearest Neighbors in C++/Python, 2018. URL: https://pypi.org/project/annoy/, python package version 1.13.0.

[29] N. Rethmeier, I. Augenstein, A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives, ACM Comput. Surv. 55 (2023) 203:1–203:17. URL: https://doi.org/10.1145/3561970. doi:10.1145/3561970.

[30] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], 2005, pp. 1473–1480. URL: https://proceedings.neurips.cc/paper/2005/hash/a7f592cef8b130a6967a90617db5681b-Abstract.html.

[31] J. Fernandez-Hernandez, L. Araujo, J. Martinez-Romo, Generation of social network user profiles and their relationship with suicidal behaviour, Procesamiento del Lenguaje Natural 72 (2024) 87–98.