MentalRiskES: A New Corpus for Early Detection of Mental Disorders in Spanish

Alba María Mármol-Romero[1], **Adrián Moreno-Muñoz**[1],
Flor Miriam Plaza-Del-Arco[2], **M. Dolores Molina-González**[1], **Arturo Montejo-Ráez**[1]
[1]Universidad de Jaén, [2]Bocconi University
[1]Campus Las Lagunillas, 23071, Jaén, Spain
[2]Via Sarfatti 25, 20100, Milan, Italy
[1]{amarmol, ammunoz, mdmolina, amontejo}@ujaen.es
[2]flor.plaza@unibocconi.it

Abstract

With mental health issues on the rise on the Web, especially among young people, there is a growing need for effective identification and intervention. In this paper, we introduce a new open-sourced corpus for the early detection of mental disorders in Spanish, focusing on eating disorders, depression, and anxiety. It consists of user messages posted on groups within the Telegram message platform and contains over 1,300 subjects with more than 45,000 messages posted in different public Telegram groups. This corpus has been manually annotated via crowdsourcing and is prepared for its use in several Natural Language Processing tasks including text classification and regression tasks. The samples in the corpus include both text and time data. To provide a benchmark for future research, we conduct experiments on text classification and regression by using state-of-the-art transformer-based models.

Keywords: Early Mental Disorder Risk Detection, Corpus, Spanish

1. Introduction

According to a recent report by the World Health Organisation (WHO), there is one in every eight people in the world suffering from a mental disorder (World Health Organization, 2022b). The COVID-19 pandemic has raised the prevalence of anxiety and depression to more than 26% in just one year. Suicide is the fourth leading cause of death among 15-29 year-olds. The organisation considers that early identification is a crucial effective intervention to prevent these problems. In Europe, the prevalence of any mental disorder among the 5–18 year-old population is 15.5% (Sacco et al., 2022). Substantial evidence suggests a noteworthy connection between excessive youth engagement with social media and adverse mental health consequences, specifically an increase in symptoms of depression and anxiety, as well as heightened levels of stress (Shannon et al., 2022).

This explains the growing interest of the scientific community in detecting and identifying mental disorders in general and, especially, in social media and messaging platforms from user messages. Natural language processing (NLP) methods offer promising results for automatically identifying profiles of people at risk or with mental health problems including depression, anxiety, and Eating Disorders (EDs). Relevant evaluation campaigns like the Cross-Lingual Evaluation Forum (CLEF) have hosted during the last years the Early-Risk Identification task (eRisk) (Parapar et al., 2021). However, most of the research conducted in this area is mainly in English. In order to foster research on the Spanish language, we release in this paper a new corpus for early mental risk detection in Spanish.[1] The corpus focuses on three major disorders: ED, depression, and anxiety. Furthermore, we offer benchmark experiments for both regression and classification tasks to facilitate further investigation in this domain and language. Additionally, this corpus has contributed to the MentalRiskES (Mármol-Romero et al., 2023) shared task at the Iberian Languages Evaluation Forum (IberLEF), held in conjunction with the Spanish Society for Natural Language Processing, as detailed by Mármol et al. (2023).

The paper's structure is as follows: Section 2 provides an overview of corpora, methods and algorithms for the early detection of mental disorders, presenting the current state-of-the-art in this field. Section 3 contextualizes the new corpus, and how it was compiled, processed, and annotated. Section 4 describes the newly created corpus and some relevant analysis on the final dataset. Then, some experiments have been performed in order to provide a baseline of some approaches to the automatic detection of mental disorders, using the corpus as a benchmark, as de-

[1] The annotation guides and datasets are available upon request in the repository: https://github.com/sinai-uja/corpusMentalRiskES

LREC-COLING 2024, pages 11204–11214
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

tailed in Section 5. In Section 6, a discussion is open on the corpus itself and these first experimental results. Finally, some conclusions and proposals for further research are provided in Section 7.

2. Related Work

According to the American Psychiatric Association (Hornberger et al., 2021), a significant number of young people experience concerns regarding their eating habits, weight, or body image. With the growing emphasis on apparent health and well-being within our culture, these issues have been exploited by social networks, as indicated by some studies (Marks et al., 2020; Aparicio-Martinez et al., 2019). Depression is estimated to occur among 1.1% of adolescents aged 10-14 years, and 2.8% of 15-19 year-old. It is estimated that 3.8% of the population suffers from depression, 5% of adults and 5.7% of those over 60 (Institute of Health Metrics and Evaluation, 2019). Social withdrawal can worsen isolation and loneliness, often associated with internet addiction (Kato et al., 2020). Finally, WHO (World Health Organization, 2022a) highlights that depression, anxiety, and behavioural disorders are leading causes of adolescent illness and disability, underscoring the need for proactive efforts to detect and address these issues.

2.1. Early Detection of Mental Disorders

In the era of deep learning, most contemporary systems have adopted an “end-to-end” methodology that, in contrast to classical approaches in computational linguistics, prioritizes a fully statistical approach (Milintsevich et al., 2023; Esackimuthu et al., 2022; Benítez-Andrades et al., 2021), eliminating the need for traditional feature engineering.

Despite the transition to end-to-end approaches, traditional techniques including bag of words, lexicons, linguistic features, and other traditional NLP representations, persist in contemporary systems alongside shallow learning models like SVM, Random Forest, and others (Espel-Huynh et al., 2021; Sadeh-Sharvit et al., 2020; Vasha et al., 2023).

A third category in this classification spectrum is hybrid approaches, which combine linguistic information with deep encodings for enhanced performance, as indicated by some research. For instance, the approach presented in Mármol-Romero et al. (2022) is based on the use of sentence embeddings from Transformers alongside features such as volumetry, lexical diversity, complexity metrics, and emotion-related scores. Also, (López Úbeda et al., 2019; López-Úbeda et al., 2021) developed a system to automatically detect anorexia in textual data. They first created a corpus of Spanish tweets covering discussions about anorexia and healthy eating habits. Their approach involved applying monolingual and multilingual transformer-based methods for tweet-level detection. For depression and anxiety detection, Burkhardt et al. (2022) evaluate the effectiveness of emotion features extracted via a BERT-based model compared to word count-based emotions as predictors.

A small number of international forums address these issues from an NLP perspective, with eRisk@CLEF being one of the most notable (Parapar et al., 2022). Nevertheless, it is worth noting that the corpora used in these tasks are in English.

2.2. Data Available

While the literature offers numerous datasets labelled for mental health risk disorders, most of them focus on English. We will primarily mention the datasets from the eRisk@CLEF Workshop,[2] which has gained considerable popularity in this field. These datasets were initially introduced in 2017 (Losada et al., 2017), specifically focusing on detecting early signs of depression and anorexia. They contain subject-generated content, such as posts and comments from social media platforms. Subjects are classified into two categories: those who suffer from the respective mental disorder and control subjects. Each subject’s dataset includes a chronological sequence of their submissions. In 2019, the organizers released an extension of these datasets that focused on early signs of self-harm (Losada et al., 2019). In 2021, they introduced a new dataset designed for identifying early signs of pathological gambling (Parapar et al., 2021; Mármol-Romero et al., 2023), which comprised subjects who experience disordered gambling and control subjects.

Unfortunately, despite Spanish being widely spoken globally, we identify very few corpora related to the mental disorders discussed in this paper. The SAD corpus (López Úbeda et al., 2019) comprises 5,707 tweets shared by individuals suffering from ED, as well as tweets from control users. While this dataset is useful for identifying a particular disorder, it does not place a focus on early detection. In related studies, like (Leis et al., 2019), linguistic features and behavioural patterns in tweets were identified as signs of depression. They used three non-public tweet datasets: one featuring the timelines of 90 users discussing their depression experiences openly, another containing manually chosen depressive tweets from these users, and a control dataset with timelines from 450 randomly selected users. Additionally, in (Pri-

[2] https://erisk.irlab.org/

eto et al., 2014), Spanish datasets were created (not publicly available), comprising 160 labelled as depression-positive and 3,093 as control based on specific sentence criteria in tweets.

3. MentalRiskES creation process

In this section, we provide an overview of the corpus, detailing its scope, curation process, compilation methodology, annotation procedures, and the agreement achieved among annotators.

3.1. Scope

As previously mentioned, we focus on three prevalent mental disorders, namely ED, depression, and anxiety. These disorders have been selected due to their significant impact on individuals' lives, the potential for public health interventions to address them effectively, and their prevalence in both social media and messaging platforms such as WhatsApp or Telegram. Anorexia nervosa is characterized by an intense fear of gaining weight, distorted body image, and extreme food restriction. Depression refers to persistent sadness, loss of interest, and impaired functioning. Anxiety involves excessive and uncontrollable worrying and various physical and psychological symptoms. The dataset comprises various subject messages exchanged within Telegram. Telegram stands out as a messaging application that facilitates the creation of thematic groups where users can discuss specific issues. This feature has allowed us to easily and legally obtain thread messages, as these are public groups where participants share their experiences. Compared to other platforms like Twitter or Instagram, where the diversity of topics is considerably broader, Telegram offers a more targeted environment that aligns optimally with our research objectives. Also, other platforms, such as Reddit, do not have much content in Spanish.

3.2. Compilation

We gathered data from publicly accessible groups on the Telegram messaging application to facilitate research aimed at early detection of mental disorders among Spanish-speaking users. Our approach involves targeting various public groups on Telegram to collect user messages related to the previously mentioned mental disorders. This data was retrieved in May 2022 and was selected to provide a comprehensive representation, taking into account both the quantity of messages and users.

In Table 1, we enumerate the names of the public groups. The group name is the actual name of the group, while the Telegram group is the identifier of the group in the platform. Notably, we gathered data from multiple groups for ED to ensure a substantial user dataset. In contrast, for the other disorders, a single group yielded a sufficiently large number of messages, making additional sources unnecessary. The datasets comprise user-generated content, with each subject's dataset featuring a chronological sequence of their comments posted in Telegram.

3.3. Curation

Following dataset compilation, we performed preprocessing, which involved removing messages with three or fewer tokens, converting emojis to text, tagging URLs, hashtags, and bold text, and anonymizing by omitting personal information such as names, aliases, and phone numbers.

On the other hand, we excluded subjects whose message count did not meet a specific amount. For ED, we applied a minimum limit of 10 messages and a maximum limit of 50. For all other conditions, the maximum limit was 100 messages. If a subject exceeded these limits, we truncated their messages to the most recent 50 or 100, depending on the specified limit.

The dataset is also provided with the emojis in their original, unprocessed format.

3.4. Annotation

In this section, we outline the process of annotating the dataset.

3.4.1. Annotation guidelines

To design the annotation guidelines for each dataset, we relied on definitions and concepts provided by major institutions such as WHO,[3] Spanish Society of Internal Medicine (SEMI)[4] and Spanish Federation of Associations for the Help and Fight against Anorexia and Bulimia Nervosa (FEACAB).[5] In addition, we provide examples of conversations from the datasets to give more clarity to each label, a list of frequently asked questions, and a graphical outline to facilitate understanding. The labels considered for each disorder were determined after examining the different subject profiles in the dataset.[6]

Following a preliminary corpus analysis, we identified the most suitable classes for each disorder. First, we skimmed through the messages and determined the labels we considered appropriate. We then did a small round test to check that these labels made sense by annotating a small set of the

[3] https://www.who.int/

[4] https://www.fesemi.org/

[5] https://feacab.org/

[6] The annotation guides can be found in the repository https://github.com/sinai-uja/corpusMentalRiskES

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

Table 1: Telegram groups used to create the corpus. ED: Eating Disorder.

data. For the datasets related to ED and depression, we used the following labels:

- **Suffer**: a user experiences everyday situations, wishes, or actions related to the ED or depression.
  - **Suffer+against**: User seeking or offering help or information to overcome the disorder. Users who foster an environment for overcoming the disorder.
  - **Suffer+in favour**: User encouraging others to explore the disorder. Users who do not foster an environment for overcoming the disorder.
  - **Suffer+other**: The user who might be suffering from the disorder does not fall into the above categories.
- **Control**: a user does not show symptoms of suffering from the disorder.

After the round test annotation, it was found that for anxiety, the annotators did not consider the different categories within the 'suffers' label to be necessary, so the following labels were used for the anxiety dataset:

- **Suffer**: a user experiences everyday situations, wishes, or anxiety-related actions.
- **Control**: a user does not show symptoms of suffering from the disorder.

We selected these labels based on observed distinctions among the positive subjects in preliminary analyses. Some exhibited a predisposition to persuade others and immerse themselves in disorder (suffer+in favour), while others primarily sought help to overcome it (suffer+against). We considered it a novel and valuable approach to detect subjects who might be suffering from the disorder in this manner in order to determine their degree of involvement in the disorder.

3.4.2. Annotation process

We used the popular platform Prolific[7] for annotating our collected data. Prolific served as the medium for recruiting human annotators, allowing us to set specific requirements such as language proficiency, education level, and age range and ensuring a balanced representation of male and female participants. In our case, we welcomed participants from any location worldwide, as long as Spanish is their first language, they hold an undergraduate degree, and their age falls between 22 and 45. We also aimed for a balanced representation of male and female participants. Annotators were provided with an annotation guide for each disorder and instructed to use the Doccano platform (Nakayama et al., 2018) for the annotation hosted on our server. Doccano is an open-source annotation platform. It was chosen for its user-friendly interface and features designed specifically for annotating datasets. Compared to other annotation platforms, Doccano offers flexibility, customization, and easy integration, making it an efficient tool for our annotation process.

To establish a connection between Prolific and Doccano, we developed a small platform that associated each Prolific user with Doccano's annotator key.

To ensure a smooth annotation process, we conducted a one-month trial task from October 21st, 2022, involving 100 subjects and ten annotators per disorder. This trial phase helped us clarify annotation guidelines, estimate labelling time, and become familiar with the platform.

After the trial phase, to complete the annotation of the users collected, we proceeded to conduct several annotation tasks in Prolific for each disorder by pooling 100 users in each of the tasks. In total, four more studies were conducted for depression, four more for anxiety, and three more for eating disorders. Each annotation task lasted approximately one and a half hours and involved

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

ten annotators[8] responsible for labelling 100 subjects. In the special case of eating disorders, one of the tasks contained only 35 subjects. Annotators who did not meet these attention check criteria were excluded. The comprehensive annotation of all datasets took approximately three months to complete.

3.4.3. Agreement between annotators

We used Fleiss's kappa (Moons and Vandervieren, 2023) and Cohen's kappa (Cohen, 1960) to measure the level of agreement between the annotators. We computed these values for each subset of the data we published and then calculated an average. The final results are in Tables 2 and 3. As expected, the agreement level decreases when we account for four labels compared to two. To provide a broader perspective, if we specifically examine the agreement on the two-label classification, it becomes evident ED poses a greater challenge for annotators to label, with anxiety and depression following as the next most challenging conditions.

Dataset	4 labels	2 labels
ED	0.185 (Slight)	0.249 (Fair)
Depression	0.316 (Fair)	0.521 (Moderate)
Anxiety	-	0.449 (Moderate)

Table 2: Cohen's kappa scores for each dataset, whether binary or multi-label.

Dataset	4 labels	2 labels
ED	0.149 (Slight)	0.223 (Fair)
Depression	0.305 (Fair)	0.521 (Moderate)
Anxiety	-	0.439 (Moderate)

Table 3: Fleiss's kappa scores for each dataset, whether binary or multi-label.

The very low score in the ED dataset is because we used two different sets of user collections, one where the majority were positive users and the other where the majority were control users. They were annotated separately, and for this reason, because there were so few users of one class and so many of the other when two people disagree, it penalizes a lot (Feinstein and Cicchetti, 1990).

After annotating the corpora and seeing the low inter-annotator agreement obtained in ED, we decided that it was more coherent to link the respective classes to the risk of a subject suffering from an ED as there were hardly any subjects for the "Suffer+other" (1 subject) and "Suffer+against" (6 subjects) classes. Fleiss's kappa and Cohen's kappa score obtained with all classes was slightly in agreement.

4. Overview of MentalRiskES

Three distinct datasets are showcased in this paper, each focusing on a specific mental health condition: ED, depression, and anxiety. The first and the last contain subjects who can be considered at risk for a disorder and those who are not, while the dataset about depression contains control subjects and subjects who might be suffering from the disorder divided into three categories (Suffer+against, Suffer+in favour, Suffer+other) to allow for multi-class classification. Each dataset contains a collection of subjects with (a) the ID of the user, (b) a chronological record of messages posted by the user in the specific Telegram group, (c) timestamps indicating when these messages were sent, and (d) the individual labels assigned by each of the ten annotators.

4.1. Corpus Analysis

The dataset for ED comprises 335 subjects (Table 4). In contrast, the depression dataset is more extensive, consisting of 449 subjects (Table 5). This higher number is attributed to the greater number of classes. Lastly, the anxiety dataset presents unbalanced data, involving 500 subjects (Table 6). All tables include the number of subjects, messages, words and vocabulary (set of different words used by a subject). In addition, the mean and variance for the last 3 items are included.

	Control	Suffer	Total
Num. subjs.	192	143	335
Num. msgs.	6,586	3,913	10,499
Mean msgs.	34.30	27.36	31.34
Std msgs.	15.24	14.58	15.33
Num. words	93,614	64,667	158,281
Mean words	487.57	452.22	472.48
Std words	575.96	660.11	612.59
Num. vocab.	46,142	32,555	78,697
Mean vocab.	240.32	227.66	234.92
Std vocab.	160.28	210.21	183.07

Table 4: Statistics for Eating Disorder (ED) per class.

In general, there is not a big difference between the number of messages in the control classes and the positive classes, although in all cases, the average number of messages is higher in the control classes. However, the same does not occur with

[8] these annotators were not consistently the same, meaning that different sets of annotators were involved in different annotation tasks

	Control	In favour	Against	Other	Total
Num. subjs.	166	240	80	13	499
Num. msgs.	6,287	7,926	2,756	401	17,370
Mean msgs.	37.87	33.03	34.45	30.85	34.81
Std msgs.	27.77	20.64	23.10	22.26	23.84
Num. words	84,433	110,926	45,101	3,839.00	244,299
Mean words	508.63	462.19	563.76	295.31	489.58
Std words	493.50	315.36	392.41	222.18	395.46
Num. vocab.	11,886	11,299	6,597	1,362	20,638
Mean vocab.	243.83	228.96	267.74	173.46	173.46
Std vocab.	171.53	115.46	133.20	102.92	139.71

Table 5: Statistics for depression per class. ‘In favour’, ‘Against’ and ‘Other’ are the subclasses of the category ‘Suffer’.

	Control	Suffer	Total
Num. subjs.	57	443	500
Num. msgs.	2,222	16,293	18,515
Mean msgs.	38.98	36.78	37.03
Std msgs.	33.22	28.16	28.76
Num. words	31,248	276,853	308,101
Mean words	548.21	624.95	616.20
Std words	388.17	513.19	500.78
Num. vocab.	14,958	126,522	141,480
Mean vocab.	262.42	285.60	282.96
Std vocab.	149.31	168.46	166.41

Table 6: Statistics for anxiety per class.

the number of words, which implies that although the positive subjects write fewer messages, they use more words in them than the control subjects, who write longer messages. There are no major differences among the three datasets as they all have similar values, with the ED dataset having the lowest average number of messages per subject. On the other hand, the dataset on anxiety is unbalanced because the nature of the Telegram group, from which the data were sourced, exhibited a lower prevalence of subjects who did not appear to be at risk of suffering from anxiety. Unlike the approach adopted for eating disorders, where an additional group was sought to balance the classes, this was not done in this case because we did not have more resources.

4.2. Examples

Tables 7, 8, and 9 exhibit the initial 5 messages from randomly selected subjects across different datasets, representing individuals with an eating disorder (ED), those at risk of depression, and those at risk of anxiety, respectively. Each subject has a score associated with each label, calculated by dividing the number of times a label appears by the total number of annotators, 10 in this case. Examples contain the emojis in text format.

Messages	Date
Este es el grupo oficial? This is the official group ?	2021-12-20 22:14:30
Me pasas el original? Can you send me the original?	2021-12-20 22:41:39
Yo estaba antes pero me sali I was here before but I left	2021-12-20 22:41:49
Alguna sabe como puedo bajar las medidas de mi brazo emoji_cara_llorando_fuerte Does anyone know how I can lower my arm measurements emoji_face_crying_loudly?	2021-12-21 18:05:00
Y de verdad lo veo muy grande And I really see it as very big	2021-12-21 18:06:36

Table 7: First 5 messages from a subject scored with a value of 0.8 as being at risk of ED. The original Spanish messages are in italics, with English translations provided below.

Messages	Date
Soy nueva cara_feliz_con_ojos_sonrientes I am new happy face_with_smiling_eyes	2021-10-11 17:11:53
Hace dos días que he salido del área de psiquiatría por intentar suicidarme ... Y no sé muy bien cómo estoy, la verdad I left the psychiatric ward two days ago after a suicide attempt... And I’m not sure how I’m doing, to be honest	2021-10-11 17:12:44
Sí, me han puesto tratamiento, pero apenas me levanto de la cama, siento que no estoy haciendo nada Yes, I have been put on treatment, but I hardly get out of bed, I feel like I’m not doing anything	2021-10-11 17:17:24
Muchas gracias a tod@s ... Thank you very much to all...	2021-10-11 18:02:25
Buenas noches gente... estoy volviendo a caer y no se como hacer... Me quiero hacer daño Good evening people... I’m falling again and I do not know how to do ... I want to hurt myself	2021-10-11 21:29:37

Table 8: First 5 messages from a subject scored with a value of 0.7 in “suffer in favour”, 0.2 in “suffer against”, 0.1 in “suffer other” and 0.0 in “control” in depression’s corpus. The original Spanish messages are in italics, with English translations provided below.

5. Baseline experiments

We conducted experiments with three different transformer-based models to establish a baseline benchmark for the different corpora.

We experimented with Spanish pre-trained models such as RoBERTa Base and RoBERTa Large, both from the MarIA project (Gutiérrez-Fandiño et al., 2021), and a multilingual pre-trained mDe-

Messages	Date
A mi me da ansiedad por ejemplo cuando tengo hambre I get anxiety for example when I am hungry.	2020-11-23 21:20:37
Cuando tengo el estómago vacío When I have an empty stomach	2020-11-23 21:20:50
Pero kodi es muy inestable But kodi is very unstable	2020-12-05 15:49:39
Yo me levante con ansiedad a full I woke up with full anxiety	2020-12-12 11:50:55
A mi me da por momentos *I get it at times*	2020-12-12 11:56:35

Table 9: First 5 messages from a subject scored with a value of 0.9 as being at risk of suffering anxiety. The original Spanish messages are in italics, with English translations provided below.

BERTa model (He et al., 2020). These models have shown good results on other Spanish tasks. RoBERTa Base,[9] RoBERTa Large[10] and mDeBERTa[11] can be found at the HuggingFace models' hub.[12]

For the systems training, we maintained the default hyperparameters. In addition, the epochs are different for each trained model due to the early stop that is set when a high value in the specified metric is reached. Details about different configurations of the models and the training process are shown in Table 10. All training and evaluations were conducted on 2x Intel(R) Xeon(R) Silver 4208 CPU @ 2.10GHz and 192 GB RAM.

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

Table 10: Baselines training details.

To address the corpus imbalance, we extracted a subset for our experiments. Table 11 displays the subjects corresponding to each label that we selected for inclusion in these subsets. We specifically selected 10 subjects for trial, used 175 subjects for training, and 150 subjects for evaluation in each of the corpora.

Regarding the pre-processing of the data, we took each subject and concatenated all their comments with tabulators. Then, we split this concatenation every 500 words (without splitting the original sentences) to build new subjects. That is, each subject gave rise to a set of sub-subjects due to the input sequence length limit of the Transformers models. For all experiments, we trained with the training set, used the trial set to perform the early stopping, and evaluated with the test set established in Table 11.

Disorder	Labels	Trial	Train	Test
ED	Suffer	5	74	64
	Control	5	101	86
Depression	Suffer + in favour	2	44	32
	Suffer + against	2	44	31
	Suffer + other	2	6	5
	Control	4	81	81
Anxiety	Suffer	5	151	122
	Control	5	24	28

Table 11: The dataset used for experimentation was a subset of the full corpus due to data imbalance. This table shows the number of subjects for each label in each subset.

5.1. Binary classification

ED, depression, and anxiety are evaluated in a binary classification setting in this case. Table 12 details the results with the classical metrics and the number of epochs used for each model, as we used early stopping to halt training when the macro-averaged F1 score reached its peak after each epoch. The first evaluation metrics used to evaluate the different models in this task are Macro-Precision, Macro-Recall and Macro-F1. We also used other metrics to measure the early detection of mental disorders shown in Table 13. These metrics have been used in the eRisk lab (Losada et al., 2017).

Disorder	Model	Epoch	P	R	F1
ED	mDeBERTa	2	0.84	0.84	0.81
	RoBERTa Large	4	0.82	0.83	0.81
	RoBERTa Base	3	0.78	0.74	0.69
Depression	mDeBERTa	2	0.79	0.69	0.64
	RoBERTa Large	4	0.76	0.72	0.70
	RoBERTa Base	3	0.74	0.66	0.61
Anxiety	mDeBERTa	5	0.78	0.68	0.71
	RoBERTa Large	4	0.84	0.65	0.69
	RoBERTa Base	3	0.84	0.59	0.60

Table 12: Results for binary classification. Precision (P). Recall (R).

Concerning results, mDeBERTa typically exhibits superior Macro-F1 scores across most tasks, but RoBERTa Large and Base may have advantages in terms of recall in some cases. The choice of model will depend on the relative importance

[9] PlanTL-GOB-ES/roberta-base-bne

[10] PlanTL-GOB-ES/roberta-large-bne

[11] microsoft/mdeberta-v3-base

[12] https://huggingface.co

Disorder	Model	ERDE5	ERDE30	ITP	speed	l-wF1
ED	mDeBERTa	0.31	0.08	5	0.92	0.75
	RoBERTa Large	0.16	0.10	2	0.98	0.79
	RoBERTa Base	0.19	0.13	2	0.98	0.72
Depression	mDeBERTa	0.30	0.15	2	0.98	0.72
	RoBERTa Large	0.29	0.16	4	0.95	0.70
	RoBERTa Base	0.34	0.18	4	0.95	0.67
Anxiety	mDeBERTa	0.38	0.14	4	0.95	0.67
	RoBERTa Large	0.33	0.12	2	0.98	0.91
	RoBERTa Base	0.30	0.14	2	0.98	0.90

Table 13: Results for binary classification in the early detection. Early Risk Detection Error (ERDE). LatencyTP (ITP). Latency-weightedF1 (l-wF1)

of precision and recall in the specific task we are addressing. Overall, all models achieve good performance in the early detection of disorders, with RoBERTa Large standing out in some metrics. In summary, results vary according to the dataset. Each model has strengths and weaknesses in detecting specific disorders in different corpora, and the choice of model will depend on the specific priorities and requirements of each task. In terms of the tasks, Transformer models appear to find ED detection the least challenging, followed by Anxiety, while Depression proves to be the most difficult to identify.

5.2. Simple regression

This is a regression problem for the tasks of ED, depression, and anxiety. These values are obtained by dividing the number of times a label appears by 10 annotators. We use the RMSE metric and Pearson correlation coefficient to evaluate the different models. Table 14 contains the results and the number of epochs in which each model was trained.

Disorder	Model	Epoch	RMSE	r
ED	mDeBERTa	6	0.23	0.87
	RoBERTa Large	7	0.20	0.89
	RoBERTa Base	8	0.18	0.91
Depression	mDeBERTa	4	0.34	0.68
	RoBERTa Large	6	0.39	0.50
	RoBERTa Base	5	0.28	0.77
Anxiety	mDeBERTa	4	0.26	0.60
	RoBERTa Large	8	0.33	0,01
	RoBERTa Base	6	0.25	0.62

Table 14: Results for simple regression. Root Mean Square Error (RMSE). Pearson correlation coefficient (r).

About the results, RoBERTa Base has a low error in the prediction and a higher value in the Pearson correlation, making it the overall superior model across all the established datasets in this scenario. Moreover, as previously mentioned in the binary classification setting, ED is easier to predict than the other tasks by the models.

5.3. Multi-class classification

Depression is evaluated in a multi-class classification setting in this case. Table 15 details the results with the classical metrics and the number of epochs used for each model. We also used metrics to measure the early detection of mental disorders shown in Table 16.

Disorder	Model	Epoch	P	R	F1
Depression	mDeBERTa	11	0.40	0.34	0.46
	RoBERTa Large	5	0.38	0.34	0.27
	RoBERTa Base	8	0.48	0.40	0.38

Table 15: Results for Multi-class classification. Precision (P). Recall (R).

Disorder	Model	ERDE5	ERDE30	ITP	speed	l-wF1
Depression	mDeBERTa	0.33	0.19	2	0.98	0.70
	RoBERTa Large	0.28	0.23	2	0.98	0.65
	RoBERTa Base	0.31	0.21	2	0.98	0.66

Table 16: Results for multi-class classification in the early detection. Early Risk Detection Error (ERDE). LatencyTP (ITP). Latency-weightedF1 (l-wF1)

In these experiments, it is clear that the results could be much better, so extensive research is needed. Although solving the multi-classification task with a more fine-grained set of labels seems to be rather difficult, the position adopted by the user (whether in favour of the disorder, fighting against it, or providing support…) is worth studying.

5.4. Multi-output regression

Depression is also evaluated in a multi-output regression setting. We use the RMSE metric and Pearson correlation coefficient to evaluate the different models. Table 17 contains the results and the number of epochs in which each model was trained.

Disorder	Model	Epoch	RMSE mean	r mean
Depression	mDeBERTa	6	0.23	0.48
	RoBERTa Large	7	0.44	-0.21
	RoBERTa Base	3	0.41	-0.15

Table 17: Results for multi-output regression. Root Mean Square Error mean of all classes (RMSE mean). Pearson correlation coefficient mean of all classes (r mean).

As in the previous task, multi-output regression is a difficult challenge. In this case, mDeBERTa is the model that obtains better results with the lowest value of error and the highest Pearson correlation. The other models still fall far short of the predictions but it would be good if a system could predict as similar as the real annotators. Although the task, again, exhibits a high degree of difficulty, the regression-based evaluation is better for evaluating how a system is close to the level of agreement of human evaluation.

6. Discussion

The corpus contains conversations of 1,334 different users over the Telegram platform, with a total number of 46,386 messages. All the users have been annotated according to the categories defined related to eating disorders, depression, and anxiety. The most unbalanced partition is that of anxiety, with only 57 control subjects in comparison to 443 users that have been annotated as "suffer". However, control subjects could be drawn from depression in order to complete a total of 223 control cases. These sizes allow for interesting research tasks.

The agreement that we found is low in general when dealing with a multi-class problem. Regarding binary annotation, it seems that is easier to agree when an eating disorder is being suffered compared to the two other disorders, depression and anxiety.

The time-stamped messages in the conversations enable early-detection analysis and evaluation. We conducted initial experiments in automatic classification and early detection, which can serve as baselines for future research. These experiments reveal varying ease of detection among the disorders, with ED being the easiest and anxiety proving to be the most challenging.

7. Conclusions and Future Work

In this paper, we present a new extensive dataset for evaluating early risk detection in three mental disorders: ED, depression, and anxiety. This dataset includes data from over 1,300 subjects and comprises more than 45,000 messages collected from various Telegram groups. The corpus will be freely available and will allow the research community to measure the performance of approaches, enabling results reproducibility and comparison. There is no similar corpus in Spanish for the early detection of mental risk and this new dataset is expected to contribute to mental health research.

Furthermore, we have performed experiments on various tasks, serving as benchmarks for future research. With message timestamps and subject-level thread labels, this dataset enables not only classification or regression studies but also early risk detection tasks, allowing us to measure the speed at which systems can identify positives over time.

We have organised the MentalRiskES (Mármol-Romero et al., 2023) evaluation campaigns using the corpus describe in this paper, and we intend to organise further evaluation forums to promote research on mental disorders, particularly within the Spanish-speaking context. Concurrently, as part of the ongoing campaign, we have undertaken cross-disorder experiments to explore whether insights gained from one disorder's context can enhance the ability of NLP methods to detect related disorders.

It is important to note that despite the intense annotation work and agreement curation, subjects have not been evaluated by experts in these disorders, so the labels should be taken not as a diagnosis of the subject but as a risk-related association. We expect to create new datasets with clinical evaluations that could serve to validate the quality of non-expert or automatically annotated corpora.

8. Ethics Statement

The main objective of this study is to contribute to society by developing a dataset that serves as a tool to create systems with scientific purposes and develop artificial intelligence to enable early detection of different mental disorders in young people on social networks. The data has been tagged by applying gender diversity and non-discrimination measures through the annotation platform. In addition, all real user identifiers were removed from the Telegram application and new identifiers were created. We insist that all results only be used for non-clinical research. People seeking help should seek it from professional psychiatrists or clinicians.

9. Acknowledgements

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, project PRECOM (SUBV-00016) funded by Spanish Ministry of Consumer Affairs, and the Big Hug project (P20_00956, PAIDI 2020), funded by the Andalusian Regional Government.

10. Bibliographical References

Pilar Aparicio-Martinez, Alberto-Jesus Perea-Moreno, María Pilar Martinez-Jimenez, María Dolores Redel-Macías, Claudia Pagliari, and Manuel Vaquero-Abellan. 2019. Social media, thin-ideal, body dissatisfaction and disordered eating attitudes: An exploratory analysis. *International journal of environmental research and public health*, 16(21):4177.

José Alberto Benítez-Andrades, José Manuel Alija-Pérez, Isaías García-Rodríguez, Carmen Benavides, Héctor Alaiz-Moretón, Rafael Pastor Vargas, and María Teresa García-Ordás. 2021. BERT Model-Based Approach for Detecting Categories of Tweets in the Field of Eating Disorders (ed). In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 586–590. IEEE.

Hannah Burkhardt, Michael Pullmann, Thomas Hull, Patricia Areán, and Trevor Cohen. 2022. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 105–115.

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

Sarika Esackimuthu, Shruthi Hariprasad, Rajalakshmi Sivanaiah, S Angel, Sakaya Milton Rajendram, and TT Mirnalinee. 2022. SSN_MLRG3@ LT-EDI-ACL2022-Depression Detection System from Social Media Text using Transformer Models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 196–199.

Hallie Espel-Huynh, Fengqing Zhang, J Graham Thomas, James F Boswell, Heather Thompson-Brenner, Adrienne S Juarascio, and Michael R Lowe. 2021. Prediction of eating disorder treatment response trajectories via machine learning does not improve performance versus a simpler regression approach. *International Journal of Eating Disorders*, 54(7):1250–1259.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish Language Models. *arXiv preprint arXiv:2107.07253*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Laurie L Hornberger, Margo A Lane, Margo Lane, Cora C Breuner, Elizabeth M Alderman, Laura K Grubb, Makia Powers, Krishna Kumari Upadhya, Stephenie B Wallace, Meredith Loveless, et al. 2021. Identification and management of eating disorders in children and adolescents. *Pediatrics*, 147(1).

Institute of Health Metrics and Evaluation. 2019. Global Health Data Exchange (GHDx). https://vizhub.healthdata.org/gbd-results/. Accessed: 2023-05-14.

Takahiro A Kato, Naotaka Shinfuku, and Masaru Tateno. 2020. Internet society, internet addiction, and pathological social withdrawal: the chicken and egg dilemma for internet addiction and hikikomori. *Current opinion in psychiatry*, 33(3):264–270.

Rosie Jean Marks, Alexander De Foe, and James Collett. 2020. The pursuit of wellness: Social media, body image and eating disorders. *Children and youth services review*, 119:105659.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-del-Arco, María Dolores Molina-González, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña-López, and Arturo Montejo-Raéz. 2023. Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish. *Procesamiento del Lenguaje Natural*, 71.

Alba María Mármol-Romero, Salud María Jiménez Zafra, Flor Miriam Plaza-del-Arco, M Dolores Molina-González, María Teresa Martín Valdivia, and Arturo Montejo-Ráez. 2022. SINAI at eRisk@ CLEF 2022: Approaching Early Detection of Gambling and Eating Disorders with Natural Language Processing. In *CLEF (Working Notes)*, pages 961–971.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):1–14.

Filip Moons and Ellen Vandervieren. 2023. Measuring agreement among several raters classifying subjects into one-or-more (hierarchical) nominal categories. a generalisation of fleiss' kappa. *arXiv preprint arXiv:2303.12502*.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)*, pages 864–887.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2022. Overview of eRisk 2022: Early Risk Prediction on the Internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 233–256. Springer.

Rosemarie Sacco, Nigel Camilleri, Judith Eberhardt, Katja Umla-Runge, and Dorothy Newbury-Birch. 2022. A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in europe. *European Child & Adolescent Psychiatry*, pages 1–18.

Shiri Sadeh-Sharvit, Ellen E Fitzsimmons-Craft, C Barr Taylor, and Elad Yom-Tov. 2020. Predicting eating disorders from internet activity. *International Journal of Eating Disorders*, 53(9):1526–1533.

Holly Shannon, Katie Bush, Paul J Villeneuve, Kim GC Hellemans, Synthia Guimond, et al. 2022. Problematic social media use in adolescents and young adults: systematic review and meta-analysis. *JMIR mental health*, 9(4):e33450.

Zannatun Nayem Vasha, Bidyut Sharma, Israt Jahan Esha, Jabir Al Nahian, and Johora Akter Polin. 2023. Depression detection in social media comments data using machine learning algorithms. *Bulletin of Electrical Engineering and Informatics*, 12(2):987–996.

World Health Organization. 2022a. Adolescent mental health. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health. Accessed: 2023-05-15.

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

11. Language Resource References

Leis, Angela and Ronzano, Francesco and Mayer, Miguel A and Furlong, Laura I and Sanz, Ferran. 2019. *Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis*. JMIR Publications Toronto, Canada.

López Úbeda, Pilar and Plaza-del-Arco, Flor Miriam and Díaz Galiano, Manuel Carlos and Ureña Lopez, L. Alfonso and Martín-Valdivia, María-Teresa. 2019. *Detecting Anorexia in Spanish Tweets*. INCOMA Ltd.

Losada, David E and Crestani, Fabio and Parapar, Javier. 2017. *eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations*. Springer.

Losada, David E and Crestani, Fabio and Parapar, Javier. 2019. *Overview of eRisk 2019 Early Risk Prediction on the Internet*. Springer.

López-Úbeda, Pilar and Plaza-del-Arco, Flor Miriam and Díaz-Galiano, Manuel Carlos and Martín-Valdivia, Maria-Teresa. 2021. *How Successful Is Transfer Learning for Detecting Anorexia on Social Media?*

Alba María Mármol-Romero, Flor Miriam Plaza-del-Arco, and Arturo Montejo-Ráez. 2023. NSINAI at eRisk@ CLEF 2023: approaching early detection of gambling with natural language processing. *Working Notes of CLEF*, pages 18–21.

Parapar, Javier and Martín-Rodilla, Patricia and Losada, David E and Crestani, Fabio. 2021. *eRisk 2021: Pathological Gambling, Self-harm and Depression Challenges*. Springer.

Prieto, Víctor M and Matos, Sergio and Alvarez, Manuel and Cacheda, Fidel and Oliveira, José Luís. 2014. *Twitter: a good place to detect health conditions*. Public Library of Science San Francisco, USA.