UMUTeam at eRisk@CLEF 2024: Fine-Tuning Transformer Models with Sentiment Features for Early Detection and Severity Measurement of Eating Disorders

Notebook for the eRisk Lab at CLEF 2024

Ronghao Pan[1], José Antonio García-Díaz[1,*], Tomás Bernal-Beltrán[1] and Rafael Valencia-García[1]

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

Abstract

This paper describes the participation of the UMUTeam in the eRisk shared task organized at CLEF 2024. We have addressed the Task 2 and 3 which are related to early detection of signs of anorexia and measuring the severity of eating disorder signs. For this purpose, several approaches were used, including the fine-tuning of a sentence transformer model for measuring the severity of eating disorder signs and the fine-tuning of pre-trained Transformers-based language models with sentiment features for detecting anorexia signs. For Task 2, we have reached the 5th position in the decision-based evaluation ranking and raking based evaluation ranking. As for Task 3, we have obtained 5th place, out of 5 participants, however, our model has a more balanced overall accuracy and performance across most metrics.

Keywords

Mental disorders, Deep learning, Natural Language Processing, Fine-tuning, Transformers

1. Introduction

Mental health is the state of a person's psychological and emotional well-being. It includes the ability to manage emotions, cope with stress, maintain satisfying relationships, work productively, and contribute to the community. It can be influenced by many factors, including genetics, life experiences, social environment, stress, and brain chemistry [1]. In recent years, there has been an increase in mental illness, an alarming phenomenon that has captured the attention of public health officials, experts, researchers, and governments around the world. According to a recent report by the World Health Organization (WHO), one in eight people in the world suffers from a mental illness[1]. Therefore, there is an urgent need to address the factors contributing to the increase in these diseases and to implement effective strategies to improve the mental and physical health of the world's population.

Several studies have shown that excessive use of social networking site can have negative effects on mental health, specially in adolescents and young adults, making it a topic of growing interest and concern in research and public health [2]. This relationship highlights the importance of early detection of mental health symptoms in order to effectively intervene and prevent these problems from worsening.

For this reason, the interest in the detection and identification of mental disorders in social network streams has grown in recent years, driven by the use of advanced Natural Language Processing (NLP) technologies, due to the increasing prevalence of mental health problems and their relationship with digital platforms [3]. In addition, a number of mental health-related tasks have emerged in important

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

✉ ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); tomas.bernalb@um.es (T. Bernal-Beltrán); valencia@um.es (R. Valencia-García)

iD 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0003-2457-1791 (R. Valencia-García)

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

evaluation campaigns, such MentalriskES [4] of Iberian Languages Evaluation Forum (IberLEF) and eRisk [5] of Conference and Labs of the Evaluation Forum (CLEF).

The eRisk Lab focuses on the development of assessment methodologies and metrics for the early detection of risks on the Internet, specially related to health and safety issues. The initiative was initiated at CLEF in Dublin in 2017, and has already hosted eight editions through 2024. Throughout these editions, the Lab has presented numerous collections and models that address different application domains. Previous editions have explored topics such as depression, eating disorders, gambling, and self-harm detection. Lab tasks include early warning and severity assessment challenges, which involve automated analysis of temporal text streams to predict specific risks and compute detailed symptom estimates from users' writings.

The eRisk@CLEF 2024 [6, 7] focuses on the early detection of signs of anorexia, the search for symptoms of depression, and measuring the severity of signs of eating disorders. This shared task was defined using the test collection, and evaluation metrics were proposed.

This paper presents the participation of the UMUTeam in tasks related to the early detection of signs of anorexia and measuring the severity of signs of eating disorders. For this purpose, several approaches have been employed, including fine-tuning of a sentence transformers model to measure the severity of the signs of eating disorders and fine-tuning of the pre-trained language models based on Transformers with sentiment features for the detection of signs of anorexia. The rest of the paper is organized as follows. Section 2 presents the task and the provided dataset. In Section 3, the methodology of our proposed system for addressing each task is described. Secondly, Section 4 shows the results obtained, and a discussion of them is presented. Finally, Section 5 concludes the paper with some conclusions and perspectives for future work.

2. Task description

This edition of eRisk focuses on detecting symptoms of depression, signs of anorexia, and the severity of symptoms associated with eating disorders through various datasets and challenges involving automated analysis of temporal text streams to predict specific problems and compute detailed symptoms estimations based on user writings. Thus, this shared task is divided into three tasks:

- **Task 1: Search for symptoms of depression**. This task is a continuation of eRisk 2023's Task 1, involves ranking sentences from user writing based on their relevance to symptoms of depression outlined in the BDI questionnaire.

- **Task 2: Early detection of signs of anorexia**. This task is a continuation of eRisk 2018's T2 and 2019's T1 tasks, focuses on early detection of signs of anorexia. In this case, we are tasked with sequentially processing pieces of evidence to detect early signs of anorexia as early as possible, primarily using Text Mining solutions on social network texts. The test collection follows the format of the collection described in [8] and comprises writings of social media users, categorized into individuals with anorexia and control users.

- **Task 3: Measuring the severity of the signs of eating disorders**. This task involves estimating the level of features associated with an eating disorder diagnosis from a history of user posts. In this task, participants are given a history of each user's posts and are asked to complete a standard eating disorder questionnaire based on the clues found in the posts. The questionnaires are derived from the Eating Disorder Examination Questionnaire (EDE-Q), which is a 28-item self-report questionnaire adapted from the Eating Disorder Examination (EDE) semi-structured interview, and only questions 1-12 and 19-28 are used.

In this edition, we participated in Task 2 and other tasks. Table 1 shows the distribution of the training dataset. We can see that the table shows various measures of the dataset, such as the number of topics, the number of submissions (posts and comments), the average number of submissions per topic, the average number of days from first to last submission, and the average number of words per submission.

Table 1
Distribution datasets from Task 2.

	2018		2019	
	Anorexia	Control	Anorexia	Control
Num. subjects	20	132	61	441
Num. submissions (posts & comments)	7 452	77 514	24 874	228 878
Avg num. of submissions per subject	372,6	587,2	407,8	556,9
Avg num. of days from first to last submission	803,3	641,5	≈ 800	≈ 650
Avg num. words per submission	41,2	20,9	37,3	20,9

For Task 3, which is a continuation of ERISK 2022 and 2023 Task 3, we used only the 2023 dataset, which has a total of 404,404 text questions.

3. Methodology

This section details the processes, techniques, and tools used for Task 2 and Task 3.

3.1. Task 2

Figure 1 shows the general architecture of our approach for Task 2. Briefly, first, we performed a preprocessing by selecting the user messages that are most relevant for anorexia identification. Second, we divided the dataset into two subsets with an 80-20 ratio: training, a subset of data that is used to train the model, and validation, a subset of data separated from the training set that is used to evaluate the model's performance during training. Third, the last hidden state of the pre-trained language models is used to obtain the text representation, and then a sentiment analysis model is used to obtain sentiment features from the texts. Finally, the last hidden state and the logits from the sentiment analysis model are concatenated to serve as input to a neural network, which is the classification head. This network includes a normalization layer (LayerNorm), a dropout layer, linear layers with Tanh as activation function, and a linear layer at the end to obtain the anorexia identification model.

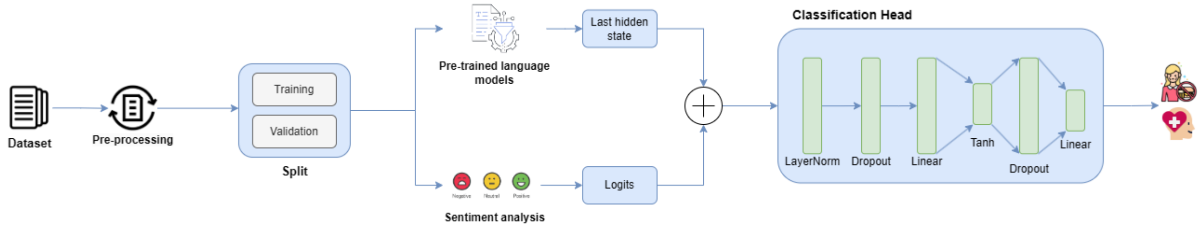

Figure 1: System architecture of Task 2.

For this task, we used only the 2019 dataset. From Table 1, we can see that at the post and comment level, there are a total of 253,752 posts, of which 24,874 are anorexia type and 228,878 are control type, indicating a significant imbalance. Therefore, we performed a preprocessing to prevent the model from always learning to predict the majority class and to reduce the noise in the dataset.

Sentiment analysis involves the use of NLP techniques to identify and categorize opinions expressed in a text, specifically to determine whether the sentiment is positive, negative, or neutral. For example, [9] shows the relationship between emotions and mental illness, as well as the importance of automatic recognition in the health field. In the context of anorexia, this analysis can help identify patterns in language that may indicate the presence of this disease. In this case, we used only negative texts from users with anorexia and positive and neutral texts from control users.

To address this task, we followed a supervised learning approach. To train our model, we used the two datasets obtained after the selection process. It is worth mentioning that the organizers only provided training data, so we selected a custom split for validation. The customized validation split is created using stratified sampling, in order to keep the balance between labels. Table 2 shows the distribution of the processed data set in the training and validation sets. We can see that we end up with a total of 4,656 texts representative of users suffering from anorexia and 11,309 of those not suffering from anorexia in the training set. In the validation set, we have a total of 1,164 anorexia type texts and 2,828 that are not related to anorexia. Moreover, we also deleted all mentions, references to URLs and hashtags from the texts, and identified and removed sequences such as "amp;format=png", "amp;s=7b66887b445eb00d7d842b15e15e15e15f4759f3deb03d", among others.

Table 2
The distribution of the training and validation split for Task 2.

	Anorexia	**No anorexia**	**Total**
Training	4 656	11 309	15 965
Validation	1 164	2 828	3 992

For this task, we evaluated the BERT [10], RoBERTa [11], and RoBERTa-large [11] models for text representation and the Cardiff NLP TweetEval model for sentiment analysis of text.

BERT [10] is a language model developed by Google in 2018 based on the Transformer architecture, a neural network designed to process data streams such as text or audio. BERT was pre-trained on large amounts of text, allowing it to capture general linguistic knowledge. This pre-trained model can then be tuned for specific natural language processing tasks such as sentiment analysis, machine translation, or question answering.

RoBERTa [11] is an extension of Facebook AI's BERT language model. It focuses on large-scale training, eliminating specific tasks and using more robust learning dynamics. These improvements make RoBERTa more effective and accurate than BERT at a variety of natural language processing tasks.

RoBERTa-large [11] is a larger and more powerful version of the RoBERTa language model. Like RoBERTa, it is based on the BERT architecture, but has more parameters and processing power. RoBERTa-large is trained on an even larger dataset for a longer period of time, allowing it to capture more complex and general linguistic patterns.

The Cardiff NLP TweetEval model [12] is a RoBERTa-based model specifically trained to perform sentiment analysis tasks on Twitter tweets. It has been trained on approximately 58 million tweets and tuned for sentiment analysis using the TweetEval benchmark dataset. Finally, for early detection, we have evaluated a strategy based on making a decision when the number of signs of anorexia in a user's messages exceeds a certain threshold.

3.2. Task 3

This task involves estimating traits associated with an eating disorder diagnosis from a set of user posts. The organizers have provided a user's posting history along with a standardized eating disorder questionnaire. Thus, the primary goal of this task is to predict potential responses to the questionnaire based on the user's posting history.

The questionnaire in question is the Eating Disorder Examination Questionnaire (EDE-Q), a 28-item self-report questionnaire derived from the semi-structured interview known as the Eating Disorder Examination (EDE). In this case, our goal is to predict responses to questions 1-12 and 19-28. The dataset consists of 28 instances of users' posting history along with their corresponding responses to the EDE-Q questionnaire.

For this task, we adopted a fine-tuning approach using a sentence transformer model that uses textual similarity to measure the similarity between potential responses (user thread text) and each question in the EDE-Q. To achieve this, we processed the user text, mapped it to the 22 questions, and assigned a score based on the user's responses to the questionnaire. To derive a scale-based score, we defined specific intervals for each possible answer within the questionnaire.

0. NO DAYS / not at all (0 to 0.1)
1. 1-5 DAYS / slightly (0.1 to 0.2)
2. 6-12 DAYS / slightly (0.2 to 0.3)
3. 13-15 DAYS / moderately (0.3 to 0.4)
4. 16-22 DAYS / moderately (0.4 to 0.5)
5. 23-27 DAYS / markedly (0.5 to 0.7)
6. EVERY DAY / markedly (0.7 to 1.0)

Thus, within the training set, each text is associated with specific questions and assigned a score, which is a randomly generated value that falls within the appropriate interval based on the user's response. We also chose a custom 80-20 split for validation. The training set contains 323,523 text-question relations along with their respective scores, while the validation set contains 80,881 such relations.

For this task, the dataset was first processed by removing contractions, mentions, hashtags, URLs, and AMP expressions, and extracting emoji features using the *emoji* Python library. Second, we fit the *multi-qa-mpnet-base-dot-v1*footnotehttps://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1 and *sentence-transformers/all-MiniLM-L6-v2*footnoteurlhttps://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 models with cosine similarity as the loss function, 10 epochs, and 1000 warm-up steps. *multi-qa-mpnet-base-dot-v1* is based on the MPNet (Multilingual Pretrained BERT) architecture, which is based on the BERT (Bidirectional Encoder Representations from Transformers) model. *sentence-transformers/all-MiniLM-L6-v2* is a kind of all-round model tuned for many use cases and trained on a large and diverse dataset of over 1 billion training pairs.

4. Results

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

4.1. Task 2

Table 3 shows the results of the different fine-tuning approaches on pre-trained language models with sentiment features in validation. We can see that the RoBERTa-base model has obtained the best performance with an M-F1 of 0.97, followed by BERT with an M-F1 of 0.95. However, RoBERTa-large, being the largest of the three models, has the worst result with an M-F1 of 0.94. Therefore, we used the RoBERTa-base fine-tuned model with sentiment features for the submissions.

Based on the summaries of the previous eRisk editions, we have seen that the DeBERTa approach has also given one of the best results. For this reason, we also evaluated the DeBERTa fine-tuning approach as the base model for our system.

For this task, we uploaded a total of 5 runs with different configurations and thresholds for the early detection approach.

- **Run 0**: This run consists of running a classification model obtained through the fine-tuning RoBERTa-base with sentiment feature within which the set of posts used has been preprocessed. The threshold used in early strategy is 10, i.e., the decision is made when more than 10 posts are identified as identifying the user's anorexia type.

- **Run 1**: This run uses the same classification model as Run 0, but uses 15 as threshold of the early detection strategy.

Table 3
Results of different fine-tuning approaches on pre-trained language models with sentiment features in the validation split of Task 2.

Model	**M-P**	**M-R**	**M-F1**
BERT	0.958116	0.949115	0.953465
RoBERTa-base	0.972686	0.969555	0.971103
RoBERTa-large	0.944243	0.951291	0.947666

- **Run 2**: In this run, we used the fine-tuned DeBERTa as a classification model and a threshold of 10 for early detection strategy.

- **Run 3**: This run uses the same classification model as Run 2, but uses a threshold of 15 for early detection strategy.

- **Run 4**: This run has the same structure as Run 2, but changing the brave strategy threshold to 20.

Table 4 shows the results of the decision-based evaluation of Task 2, specifically the precision, recall, and F1 score over the five runs. Accuracy ranges from 0.14 to 0.16, indicating a low variability in the model's ability to correctly identify relevant instances. Recall is very high across all runs, between 0.98 and 0.99, demonstrating the model's effectiveness in capturing almost all relevant instances. The F1 score, which balances precision and recall, shows a slight improvement from 0.25 in run 0 to 0.27 in run 4. The ERDE_5 and ERDE_{50} metrics, which measure early risk detection errors, remain relatively stable, indicating consistent early detection performance across all runs. Latency, which reflects the time it takes to make a correct prediction, increases from 18.0 in Run 0 to 35.5 in Run 4. Speed, which reflects the speed of processing, decreases slightly to a low of 0.87 in Run 4.

Overall, Run 4 achieves the highest accuracy and F1 score at the cost of higher latency, while Runs 0 and 2 offer lower latency at slightly lower accuracy and F1 score. With this result, we ranked fifth in decision-based evaluation.

Table 4
Results of UMUTeam for Task 2 in decision-based evaluation including the precision (P), recall (R), and F1-score (F1). Other metrics considering the performance of the methods are also included.

Run	*P*	*R*	*F1*	*$ERDE_5$*	*$ERDE_{50}$*	*$latency_{tp}$*	*speed*	*$latency_w F1$*
0	0.14	0.99	0.25	0.20	0.09	18.0	0.93	0.23
1	0.15	0.99	0.26	0.19	0.09	27.0	0.90	0.24
2	0.14	0.99	0.25	0.20	0.09	19.0	0.93	0.23
3	0.15	0.99	0.27	0.19	0.09	28.0	0.90	0.24
4	0.16	0.98	0.27	0.19	0.10	35.5	0.87	0.23

Table 5 shows the ranking based evaluation results (only 1 writing result is reported) and we can see that the five runs are identical: P@10 is systematically at 0.20, NDCG@10 at 0.12, and NDCG@100 at 0.14. The P@10 metric, which measures the accuracy in the top 10 positions, remains at 0.20. Both NDCG@10 and NDCG@100, which evaluate the quality of ranking by considering the relevance of documents in different positions, show consistent values, indicating that the model maintains a similar level of efficiency in ranking relevant results in the top 10 and top 100 positions. Overall, these results demonstrate reliable performance in ranking-based evaluation, and we have achieved 5 best results.

4.2. Task 3

For this task, we presented two runs based on fine-tuning a pre-trained sentence transformer model, that uses textual similarity to measure the similarity between potential responses (user thread text) and each question in the EDE-Q.

Table 5
Results of UMUTeam for Task 2 in ranking-based evaluation (only 1 writing result reported).

Run	P@10	NDCG@10	NDCG@100
0	0.20	0.12	0.14
1	0.20	0.12	0.14
2	0.20	0.12	0.14
3	0.20	0.12	0.14
4	0.20	0.12	0.14

- **Run 0**: This run consists of using the *multi-qa-mpnet-base-dot-v1* fine-tuned model as a system model to identify the similarity between the EDE-Q question and the user posts. For each user post, it is fed into the system and the system calculates the degree of similarity between the EDE-Q question and the post. Based on the score obtained by the system, a possible answer to the question is assigned within the intervals defined in section X. Once all the contributions have passed through the system, the most repeated answer for each question is assigned as the final answer.

- **Run 1**: This run uses the same approach as Run 0, but uses *sentence-transformers/all-MiniLM-L6-v2* fine-tuned model as a system model.

Table 6 shows the results obtained in the evaluation of Task 3, evaluated according to different metrics: MAE (Mean Absolute Error), MZOE (Mean Zero-One Error), MAE_macro, GED (Global Eating Disorder Score), RS (Restraint Subscale), ECS (Eating Concern Subscale), SCS (Shape Concern Subscale), and WCS (Weight Concern Subscale).

First, we looked at the Mean Absolute Error (MAE), which measures the average size of the errors in the predictions without considering their direction. The Run 1 achieved an MAE of 2.227, while the Run 0 achieved an MAE of 2.366. This indicates that Run 1 had a higher overall accuracy in its predictions.

The MZOE metric shows the average of the errors in terms of binary hits and misses. Run 0 had an MZOE of 0.798 compared to 0.859 for Run 1. This means that Run 0 made fewer errors and was more accurate in correctly classifying cases.

As for MAE_macro, which evaluates the mean absolute error balanced across classes, Run 1 performed better with a value of 2.286 compared to 2.833 for Run 0. This result indicates that Run 1 achieved a more balanced performance between the different data categories, which is crucial in situations where all classes are equally important.

The GED measures the overall accuracy of the model in predicting eating disorders. The Run 0 had a GED of 3.261, while the Run 1 had a GED of 3.286. Although the difference is small, Run 0 showed slightly better performance on this overall measure.

For the RS, which measures accuracy in predicting dietary restraint behavior, both runs showed very similar results, with Run 1 scoring an RS of 3.269 and Run 0 scoring 3.285. This parity indicates that both runs are comparable in terms of accuracy on this specific subscale.

On the ECS, Run 0 showed better performance with an ECS of 2.659 compared to 2.911 for Run 1. This result suggests that Run 0 was more effective at capturing specific food concerns.

On the SCS, Run 1 performed better with an SCS of 2.560 compared to 2.771 for Run 0. This data suggests that Run 1 was more accurate in predicting body shape concerns.

Finally, on the WCS, Run 1 also outperformed Run 0 with a WCS of 2.026 compared to 2.218. This demonstrates a better ability of Run 1 to predict weight concern.

In summary, although Run 1 showed better overall accuracy and more balanced performance on most metrics, Run 0 excelled in specific aspects such as MZOE, GED, and ECS.

Table 6
Results of UMUTeam for Task 3 in performance results.

Run	MAE	MZOE	$\text{MAE}_{\text{macro}}$	GED	RS	ECS	SCS	WCS
0	2.366	0.798	2.833	3.261	3.285	2.659	2.771	2.218
1	2.227	0.859	2.286	2.326	2.911	2.142	2.560	2.026

5. Conclusion

This paper summarizes UMUTeam's participation in the eRisk collaborative task of the 2024 edition of CLEF. The eRisk Lab focuses on the development of assessment methods and metrics for the early detection of risks on the Internet, especially related to health and safety issues. In this edition, the focus is on detecting symptoms of depression, early detection of signs of anorexia, and measuring the severity of signs of eating disorders in three related subtasks.

In this shared task, we have focused on Task 2 and Task 3, which are related to early detection of signs of anorexia and measuring the severity of eating disorder signs. For this purpose, several approaches were used, including the fine-tuning of a sentence transformer model for measuring the severity of eating disorder signs and the fine-tuning of pre-trained Transformers-based language models with sentiment features for detecting anorexia signs.

In Task 2, we present 5 runs based on different settings, using different fine-tuned models as the classification model for the system and different thresholds for the early detection strategy. We ranked fifth in the decision-based evaluation, and run 4 achieved the highest accuracy and F1 score at the cost of higher latency, while runs 0 and 2 offer lower latency with slightly lower accuracy and F1 score. For the decision-based evaluation, we obtained the top 5 results. In this case, all five runs are identical: P@10 is consistently 0.20, NDCG@10 is 0.12, and NDCG@100 is 0.14.

From the results obtained, we can see that the sheer number of comments may not be enough; the context and severity of the comments are also important. We also found that removing certain negative comments from users labeled as "control" runs the risk of the model not learning to properly distinguish between negative comments that are normal and those that are indicative of a mental disorder, which could degrade the performance of the system.

In Task 3, we present two runs based on fine-tuning a pre-trained sentence transformer model that uses textual similarity to measure the similarity between possible answers (user thread text) and each question in the EDE-Q. In this case, run 1, which is based on a fine-tuned model of *sentence-transformers/all-MiniLM-L6-v2*, has the best result in overall accuracy and a more balanced performance on most metrics.

As a future line, we suggest adding the user's previous context as an input to improve performance, and not removing all negative comments from users marked as "control", to avoid that the model does not learn to correctly distinguish between negative comments that are normal and those that are indicative of a mental disorder. Furthermore, it is important to examine the relationship between indicators of mental illness and hate speech [13], the use of humor [14], and the demographic and psychographic characteristics of the message authors [15].

Acknowledgments

This work is part of the research projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe and LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR, and "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the “Plan de Recuperación, Transformación y Resiliencia (PRTR)”.

References

[1] S. Dattani, L. Rodés-Guirao, H. Ritchie, M. Roser, Mental health, Our world in data (2023).

[2] R. Sacco, N. Camilleri, J. Eberhardt, K. Umla-Runge, D. Newbury-Birch, A systematic review and meta-analysis on the prevalence of mental disorders among children and adolescents in Europe, European Child & Adolescent Psychiatry (2022). doi:10.1007/s00787-022-02131-2.

[3] R. A. Calvo, D. N. Milne, M. S. Hussain, H. Christensen, Natural language processing in mental health applications using non-clinical texts†, Natural Language Engineering 23 (2017) 649–685. URL: https://api.semanticscholar.org/CorpusID:17828909.

[4] A. M. M.-R. y Adrián Moreno-Muñoz y Flor Miriam Plaza-del-Arco y María Dolores Molina-González y Maria Teresa Martín-Valdivia y Luis Alfonso Ureña-López y Arturo Montejo-Raéz, Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023) 329–350. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early Risk Prediction on the Internet, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 294–315.

[6] J. Parapar, P. Martín Rodilla, D. Losada, F. Crestani, Overview of eRisk 2024: Early Risk Prediction on the Internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024, 2024.

[7] J. Parapar, P. Martín Rodilla, D. Losada, F. Crestani, Overview of eRisk 2024: Early Risk Prediction on the Internet (Extended Overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, 2024.

[8] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.

[9] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in Spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024) e1992. doi:10.7717/peerj-cs.1992.

[10] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[12] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, TweetEval: Unified benchmark and comparative evaluation for tweet classification, arXiv preprint arXiv:2010.12421 (2020).

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

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

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