UMUTeam at MentalRiskES@IberLEF 2024: Using the Fine-Tuning Approach of Transformer-Based Models with Sentiment Feature for Early Detection of Mental Disorders

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

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

Abstract

The alarming rise in mental disorders has sparked interest in early detection through social networking. The relationship between excessive use of social media and mental health problems, especially among adolescents, has led to a growing interest in early detection of these problems through social media comments. The MentalRiskES task in IberLEF 2024 focuses on this early detection of risks of mental disorders through comments in Spanish. This paper presents UMUTeam's contribution, focusing on disease and context detection. We use pre-trained linguistic models with outputs from emotion and sentiment models. In Task 1, we ranked 15th in decision and latency based classification; in Task 2, we ranked 10th in decision and latency based classification.

Keywords

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

1. Introduction

The increase in mental illness in recent years is 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. There is no single cause for this increase, but rather a complex interplay of environmental, social and biological factors. For example, in the COVID-19 era, the prevalence of anxiety and depression increased by more than 26% in just one year. Suicide has become the fourth leading cause of death among young people aged 15-29. This situation underscores the urgency of addressing the factors contributing to the increase in these diseases and implementing effective strategies to improve the mental and physical health of the global population [1].

Much evidence shows or suggests that there is a relationship between excessive use of social networking sites by young people and their negative mental health outcomes, particularly an increase in symptoms of depression and anxiety, as well as levels of stress. This relationship highlights the importance of early identification of these symptoms in order to effectively intervene and prevent these problems before they worsen. In other words, early identification of signs of deteriorating mental health may enable parents, educators, and health professionals to take appropriate action to mitigate the negative effects [2].

For this reason, in recent years there has been a growing interest in detecting and identifying mental disorders in social network, due to the increasing prevalence of mental health problems and their relationship with the use of digital platforms. Various mental health related tasks have also emerged, such as eRisk [3] in the Cross-Lingual Evaluation Forum (CLEF) assessment campaign. However, these campaigns have mainly focused on English, leaving aside other languages such as Spanish. Therefore, the MentalRiskES [4] task in IberLEF 2024 [5] is the second edition that aims at the early detection of risks of mental disorders through comments in Spanish from social network sources. In this edition, the organizers have mainly proposed three tasks that focus on the identification of different mental

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

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

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

illnesses from different perspectives with a new corpus [6]. These tasks include disease detection, context detection, and suicidal ideation detection.

This paper presents the UMUTeam's contribution to the first two tasks, focusing on disease and context detection, based on the fine-tuning of different pre-trained Transformer-based linguistic models mixed with the outputs (logits) of the emotion and sentiment identification models with different early detection methods. The rest of the paper is organized as follows. Section 2 presents the task and the provided dataset. Section 3 describes the methodology of our proposed system to address subtask 1 and subtask 2. Section 4 presents the obtained results. Finally, section 5 concludes the paper with some conclusions and possible future work.

2. Task description

MetalRisk focuses on the early detection of mental illness through comments posted by different users on Telegram. Thus, given a history of a user's messages, the goal is to identify whether the user is suffering from a disorder and the context that influences the mental health problem. The organizers have considered the mental health detection from three perspectives: i) disorder detection, which is a multi-class classification problem whose goal is to detect whether the user suffers from depression, anxiety, or no detected disorder; ii) context detection, which is similar to the previous approach, but with the addition of identifying the context from which the mental health problem appears to originate, if a disorder is detected, and iii) suicide ideation detection, which is a binary detection problem whose goal is to determine whether the user is exhibiting symptoms of potential suicidal ideation. Therefore, one task has been proposed for each approach.

Task 1, which is a multi-class classification problem, has 3 possible labels: depression, which is characterized by persistent sadness, low mood, and lack of interest or pleasure in previously rewarding and enjoyable activities; anxiety; and none (no disorder detected). In contrast, Task 2 has the same objective as Task 1, but in this case, a multi-class classification problem is added, which consists of identifying the context from which the detected mental health problem appears to originate. In this case, the available contexts are: addiction context as "addiction", emergency context as "emergency", family context as "family", work context as "work", social context as "social" and other contexts as "other". If no context is detected, "none" is assigned. Contexts are necessary only required if the subject is predicted to have depression or anxiety.

Table 1 shows the distribution of the dataset at the user level and at the message level after preprocessing. At the user level, we can see that there are a total of 465 message histories from different users, of which 213 have no mental illness, 164 have depression, and 99 have anxiety. To build a model for identifying mental illness, we preprocessed each user's history and retained only the negative comments from users suffering from any mental illness. For those labeled "none", we removed the negative comments, leaving only the positive and neutral comments. In this way, we achieved noise reduction, clarity in mental illness patterns, and simplicity in classification. However, by eliminating negative comments from users labeled as "none", we run the risk that the model will not learn to properly distinguish between negative comments that are normal and those that are indicative of a mental disorder. In Table 1, we can see the distribution of the message-level dataset after preprocessing. There are about 9504 messages in total, of which 5931 are of the type "none", 2322 are of the type "depression", and 1251 are of the type "anxiety". The Table 2 shows the distribution of the data sets for Task 2. In this case, the problem is of the multi-label type, which means that each user suffering from a disease can be associated with more than one context.

3. Methodology

Figure 1 shows the general architecture for Task 1, which consists of early detection of mental illness using pre-trained language models, sentiment identification models, and classification techniques. Each step of the process is described below:

Table 1
Distribution of the datasets of the Task 1.

Total	None	Depression	Anxiety
User level			
465	213	164	88
Message level			
9,504	5,931	2,322	1,251

Table 2
Distribution of the datasets of the Task 2.

Addiction	Emergency	Family	Work	Social	Other	none
User level						
12	17	61	17	88	56	74

- **Dataset**: The process starts with a filtered dataset mentioned in Section 2.
- **Preprocessing**: Textual data is preprocessed to clean and prepare the text for analysis. In this case, all emoticons, hashtags, links and special characters are removed.
- **Split**: Once the dataset is preprocessed, it is split into two subsets: one for training and one for validation. This allows the effectiveness of the model to be evaluated on data not seen during training.
- **Pretrained language models**: Pre-trained language models (e.g. BERT, RoBERTa, etc.) that have already been trained on large amounts of text are used. These models generate vector representations (embeddings) of the input text.
- **Last Hidden State and Logits**: The last hidden state of the pre-trained language model is extracted, providing a deep representation of the text. Logits are the output of the *Pysentimiento* model [7], which are used to classify the text into different emotional categories, such as negative, neutral, or positive.
- **Sum of last hidden state and logits**: The last hidden state representations and the logits are combined to consolidate the information extracted from the text.
- **Classification Head**: Finally, a classification head is added, which is a combination of several layers such as LayerNorm, Dropout and Linear, and Tanh as the activation function.

Finally, the model is trained and the final output of the model is a prediction about the mental state of the text, such as depression, anxiety, or none. This architecture makes it possible to detect early signs of mental illness through text analysis, which can be crucial for early intervention and support of affected individuals.

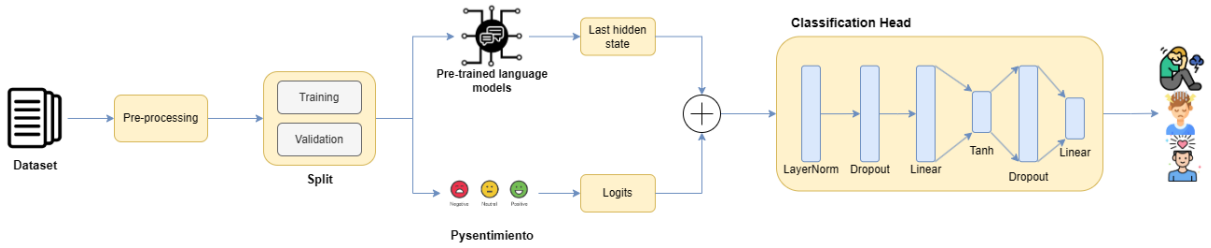

Figure 1: Overall system architecture for Task 1.

For Task 2, which aims to identify the context of mental illness, we used the same approach as in Task 1, as shown in Figure 2. However, in this case, the logits obtained from Model 1 are also added

in order to improve the performance of the model. Note that since this is a multi-label classification problem, each user may have more than one context. Therefore, in the preprocessing, we converted the labels into a one-hot representation.

The models evaluated for both tasks are: MarIA, BETO and BERTIN. All three models are of the monolingual type, i.e. they are pre-trained with a large dataset in Spanish. MarIA [8] is a transformer-based language model for Spanish. It is based on the RoBERTa base model and has been pre-trained on the largest Spanish corpus known to date, with a total of 570 GB of clean and deduplicated text. BETO [9] is a model based on BERT and pre-trained on a Spanish corpus. It is similar in size to a BERT base and has been trained using the Whole Word Masking technique. BERTIN [10] is a RoBERTa-based model and was trained from scratch on the Spanish part of mC4 using Flax [1]. The hyperparameters used to the train the model for both tasks are: 16 training batch size, 15 epochs, 0.01 weight decay and 2e-5 learning rate.

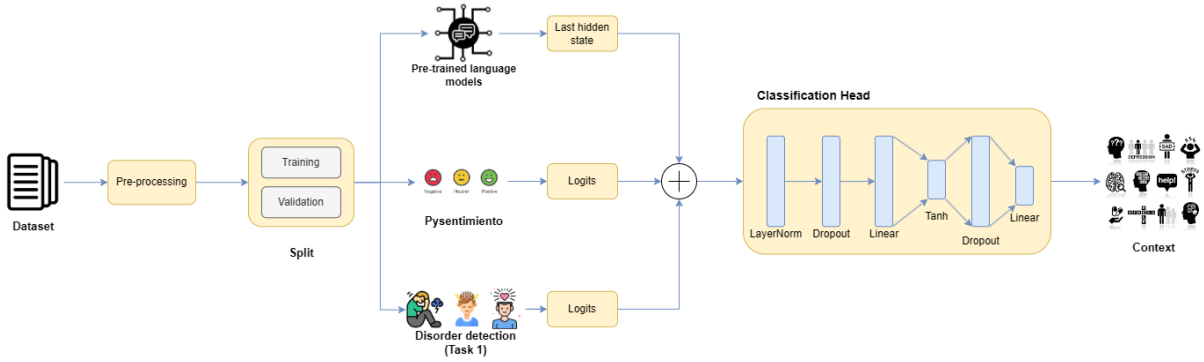

Figure 2: Overall system architecture for Task 2.

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 1

For Task 1, different pre-trained models such as BETO, MarIA and BERTIN were evaluated from two perspectives: normal fine-tuning for mental illness classification and fine-tuning of the pre-trained models by adding sentiment features. In Table 3, we can observe the results where MarIA is the most robust model in both configurations, with and without sentiment features, showing significant improvements in accuracy and F1-score when sentiment features are added. MarIA obtained an M-F1 of 76.19 in the configuration without sentiment features and 81.12 with sentiment features. In contrast, BETO and BERTIN also improve with sentiment features, but to a lesser extent than MarIA. In particular, BERTIN shows a remarkable performance recovery when sentiment features are considered. Therefore, based on the results obtained, we can conclude that the inclusion of sentiment can improve the performance of models in the classification of mental disorders.

Figure 3 shows the confusion matrix of the MarIA in the validation set. The confusion matrix shows the percentages of correct and incorrect classifications for three categories: *anxiety*, *depression*, and *none*. This analysis is important to evaluate the performance of a classification model. Overall, our model shows high accuracy in the *none* category, with 98.15% of the instances correctly classified. This indicates that the model is very effective in correctly identifying cases where neither anxiety nor depression is present. For the *depression* category, the model also shows good accuracy, with

[1] https://github.com/google/flax

Table 3
Result of different pre-trained language model in validation split of Task 1.

Model	M-P	M-R	M-F1
Without sentiment feature			
MarIA	77.2504	75.5207	76.1925
BETO	74.7810	75.5516	75.1033
BERTIN	74.3958	73.1138	73.6917
With sentiment feature			
MarIA	**81.1171**	**76.3064**	**78.1785**
BETO	76.7196	75.4402	75.5035
BERTIN	76.5461	77.1464	76.7737

76.77% of instances correctly identified. However, there is significant confusion between depression and anxiety, with 9.46% of depression cases incorrectly classified as anxiety and 13.76% classified as *none*. The category *anxiety* has the lowest performance, with 54.00% of the instances correctly classified. This suggests that the model has significant difficulty in accurately identifying *anxiety*. A significant proportion of *anxiety* cases (36.40%) are misclassified as depression.

In summary, the model performs well in identifying cases where there is neither *anxiety* nor *depression*, and performs well in identifying *depression*. However, it shows significant difficulties in distinguishing between *anxiety* and *depression*, as well as in correctly identifying cases of *anxiety*. This problem is partly due to the smaller number of *anxiety* examples in the training set. The lack of *anxiety* examples in the training data may lead to a bias in the model, making it less competent in recognizing this condition compared to the other categories.

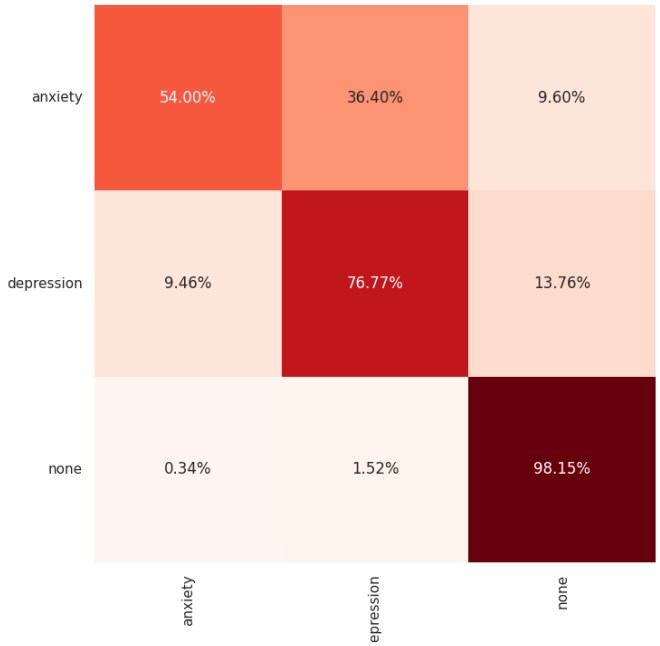

Figure 3: Confusion matrix of MarIA in validation split.

However, the models obtained operate at the sentence level, so predicting whether a user has a mental disorder requires taking into account predictions from the user's previous comments. A user's single comment about a mental disorder does not necessarily mean that they have that disorder. Therefore,

to determine whether a user is suffering from depression or anxiety, our system processes the set of user messages and uses the most common label to make the decision. For an early detection approach, we tested a set of thresholds. For example, if the threshold is set to 5, and the first 5 user comments are related to depression or anxiety, the system will conclude that the user is likely suffering from the disorder.

For this task, we submitted three runs, each with the same structure but differing in minor aspects of the configuration of the early detection method.

- **Run 0**. This run uses the MarIA model with sentiment features as the classification model and uses a threshold of 5 for early detection. This means that during each round the labels of the previous rounds are checked. If the user has 5 or more comments related to depression or anxiety, the system considers it as depression or anxiety.
- **Run 1**. This run uses the same approach as for run 1, in this case, a threshold of 10.
- **Run 2**. This run takes the longest to decide because it consists of predicting all rounds and making a final decision based on the most repeated label.

Table 4 shows the results obtained in the official ranking based on decisions. We can see that the conservative strategy of run 2 has obtained the best result with an M-F1 of 67.5, reaching the 15th place. In this method, the system makes the decision after predicting all the rounds. On the other hand, the threshold of 5 (run 0) for the early detection strategy has obtained an M-F1 of 64, reaching the 16th position. On the other hand, the threshold of 10 has obtained the worst result with an M-F1 of 26.9. Moreover, We can see that in this case, Run 2 is still the one that has obtained the best result in ERDE30 with a value of 0.166, followed by Run 0 and Run 1 with values of 0.194 and 0.501 respectively.

Table 4
Results of UMUTeam for Task 1. For each run, the rank (#), accuracy (A), macro precision (M-P), macro recall (M-R), and macro (M-F1) are reported for decision-based evaluation and ERDE5, ERDE30, latencyTP, speed, latency-weighted F1 (LWF1) are reported for latency-based evaluation.

#	Run	A	M-P	M-R	M-F1	ERDE5	ERDE30	latencyTP	speed	LWF1
16	0	63.0	**72.8**	66.2	64.0	0.593	0.194	11	0.844	0.629
28	1	51.5	71.2	35.5	26.9	0.501	0.501	80	0.154	0.013
15	**2**	**69.0**	70.1	**68.3**	**67.5**	**0.203**	**0.166**	**1**	**1**	**0.780**

4.2. Task 2

For Task 2, which aims to identify the context of users suffering from mental illness, we followed the same methodology as in Task 1. We analyzed different pre-trained Spanish language models, such as BETO, MarIA and BERTIN, for multi-label context classification. We fine-tuned these models with and without sentiment features and the logits of the mental illness classification model (MarIA model of Task 1). Table 5 shows the results obtained.

We can see that BETO stands out as the best model without the additional features, obtaining the highest mean F1 score (28.2869), suggesting that it handles the multi-label context classification better compared to MarIA and BERTIN in this configuration. The inclusion of sentiment and mental disorder features does not always improve performance. In the case of MarIA, although recall improves slightly, accuracy decreases significantly, suggesting a trade-off between these metrics. BERTIN shows an improvement in the mean F1 score when additional features are added, with an M-F1 of 26.8134 vs. 25.0449, indicating that it can benefit from this additional information, although not as much as BETO without these features.

For this task, we present three runs based on those of Task 1. That is, when the system detects that a user has a mental illness, it searches for the most repeated contexts from the previous rounds in order to have a set of contexts related to the illness.

Table 5
Result of different pre-trained language model in validation split of Task 2.

Model	M-P	M-R	M-F1
Without sentiment+mental_disorder feature			
MarIA	49.0431	20.7012	23.8928
BETO	**45.2212**	**24.5992**	**28.2869**
BERTIN	47.1883	22.6518	25.0449
With sentiment+mental_disorder feature			
MarIA	40.8600	21.6143	25.7859
BETO	35.2956	24.2577	27.0830
BERTIN	44.0429	23.9276	26.8134

Table 6 shows the BETO ranking report on the Task 2 validation partition. In this case, our model shows uneven performance in different categories. The model performs well in the Social category, but shows significant difficulty in Addiction and Emergency, with particularly low recall. The micro and macro averages indicate limited overall performance, reflecting the need for improvement in correctly classifying different classes. Overall, the model needs adjustments, such as more balanced data collection, hyperparameter tuning, and improved preprocessing techniques, such as pre-identification of mental illness, to improve its performance in identifying all categories.

Table 6
Classification report of BETO in validation split of Task 2.

	Precision	Recall	F1-score
Addiction	25.0000	13.9535	17.9104
Emergency	80.0000	6.3492	11.7647
Family	37.5000	42.8571	40.0000
None	29.2208	26.9461	28.0374
Other	46.1538	24.0000	31.5789
Social	53.2189	45.5882	49.1089
Work	45.4545	12.5000	19.6078
Micro avg	41.8733	32.2718	36.4508
Macro avg	45.2212	24.5992	28.2869
Weighted avg	44.7882	32.2718	35.1824

Table 7 shows the results obtained in the official ranking based on decisions. We can see that Run 2 has the highest accuracy with a value of 0.077 and the best macro precision (M-P) of 0.224, showing a balance between accuracy and recall, although with low absolute values in all metrics. However, it is followed by Run 0, with the best M-F1 of 22.4 and Run 1 with M-F1 of 4.4, respectively. On the other hand, Run 2 is clearly the top performer in this latency-based evaluation. It has the lowest ERDE5 (0.203) and ERDE30 (0.166) values, indicating better performance in terms of errors relative to early detection.

In the decision-based evaluation, Run 2 shows a better balance between precision and recall, although with low absolute values. In the latency-based evaluation, Run 2 excels in all key metrics, showing to be the best model in terms of early detection efficiency and accuracy.

From the results obtained, we can see that the simple number of comments may not be enough; the context and severity of the comments are also important. In this case, a threshold of 5 is better than 10, but the prediction is still more robust in all rounds as in Task 1. We have also found that removing certain negative comments from users marked as “none” 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.

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

#	Run	A	M-P	M-R	M-F1	ERDE5	ERDE30	latencyTP	speed	LWF1
11	0	.7	16.6	**40.8**	**224**	0.593	0.194	11	0.844	0.629
19	1	0	16.9	2.6	4.4	0.501	0.501	80	0.154	0.013
10	2	**.77**	**22.4**	17.0	17.8	**0.203**	**0.166**	**1**	**1**	**0.780**

In terms of carbon emissions, our approach has a mean duration of 3.156 with a deviation of 2.647 and a mean emission of 5.87e-5 with a deviation of 5.11e-5 across all runs in Task 1 and Task 2, because submissions for both tasks are uploaded at the same time. In this case, our mean duration is at the lowest of the rankings, but our mean emission is in the middle of the rankings.

5. Conclusion

This article summarizes UMUTeam's participation in the MentalRiskES shared task at IberLEF 2024. This task focuses on the early detection of mental illness from three perspectives: disease detection, context detection, and suicidal ideation detection.

In this shared task, we participated in Task 1 and Task 2, which focus on disease detection and context detection, respectively. In both tasks, we have evaluated the normal tuning approach of different pre-trained language models and also the tuning of these models with sentiment features obtained with *pysentimiento* for Task 1. In addition, we evaluated the concatenation of pre-trained language models with sentiment features and the classification model output of Task 1.

In Task 1, we achieved the 15th position with an M-F1 of 0.675 in run 2 with an ERDE30 value of 0.162. On the other hand, we obtained better results in Task 2, reaching the 10th position in the decision-based ranking with an M-F1 of 0.203 in Run 2 and an ERDE30 value of 0.166.

The results indicate that the simple number of comments may not be sufficient; it is also important to consider the context and severity of the comments. In this case, a threshold of 5 is preferable to 10, although the prediction remains more robust in all rounds. We have also found that removing certain negative comments from users marked as "None" may prevent the model from learning to properly distinguish between normal negative comments and those that indicate a mental disorder.

As a future line, we propose to incorporate context prior to current comment prediction and to test different early detection methods. We also propose to evaluate different multilingual pre-trained models based on Transformers. In addition, we think that it is relevant to consider the relationship between signs of depression and hate speech [11] [12], the use of humour [13], and the demographic and psychographic traits of the authors of the messages [14].

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] 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.

[2] H. Shannon, K. Bush, P. J. Villeneuve, K. G. Hellemans, S. Guimond, Problematic social media use in adolescents and young adults: Systematic review and meta-analysis, JMIR Ment Health 9 (2022) e33450. URL: https://mental.jmir.org/2022/4/e33450. doi:10.2196/33450.

[3] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 294–315.

[4] 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).

[5] 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.

[6] 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.

[7] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: a python toolkit for opinion mining and social nlp tasks, arXiv preprint arXiv:2106.09462 (2021).

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

[9] 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.

[10] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[11] 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.

[12] 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.

[13] 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.

[14] 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.