Mental Disorder Detection in Spanish: Hands on Skewed Class Distribution to Leverage Training

Xabier Larrayoz[1], Arantza Casillas[1], Maite Oronoz[1] and Alicia Pérez[1]

[1]*HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Donostia-San Sebastián, Spain*

Abstract

The early detection of mental disorders in online environments is becoming crucial in contemporary mental health settings. In this study, we addressed this issue by participating in the MentalRiskES competition. Supervised classification approaches are trained by means of annotated sets of instances. In this case, each instance is a message and, ideally, we would expect annotations at message level, however, annotations are given at user level. Our key contribution consisted of an innovative message re-labeling approach to enhance the inference of supervised models. Using labeled user data alongside their messages in Spanish, we evaluated various language models, including Sentence Transformers and BETO.

Keywords

Early risk prediction, Natural Language Processing, Deep learning, Mental health

1. Introduction

Mental health is an essential component of human well-being, yet its significance is often overlooked or underestimated. However, mental disorders affect millions of people worldwide, with consequences ranging from decreased quality of life to the risk of suicide. According to the World Health Organization (WHO), approximately one in eight people globally suffers from some form of mental disorder, and this figure could be increasing due to factors such as the COVID-19 pandemic [1].

The rise in anxiety, depression, and other mental disorders during the pandemic has underscored the urgent need to identify and address these issues early and effectively. Early detection of mental disorders can facilitate more effective interventions, improve treatment outcomes, and ultimately save lives.

In response to this need, the MentalRiskES competition [2, 3] has become a focal point for research in early detection of mental disorder risks in the Spanish-speaking context. This competition, now in its 2024 edition, provides a platform for researchers worldwide to develop and evaluate methods for detecting mental disorders using social media data in Spanish.

In this edition of the competition, we have focused on Task 1: disorder detection, which involves multi-class classification to determine whether a user suffers from depression, anxiety, or shows no detected disorder. This challenge opens a path for identifying language patterns associated with these disorders in online environments, which can provide early risk signals and intervention opportunities.

This paper describes our approach to address Task 1 of the competition, including our methodology, results, and discussion of key findings. Through this work, we hope to contribute to the development of more effective tools and techniques for early detection of mental disorders in the Spanish-speaking community.

IberLEF 2024, September 2024, Valladolid, Spain

xabier.larrayoz@ehu.eus (X. Larrayoz); arantza.casillas@ehu.eus (A. Casillas); maite.oronoz@ehu.eus (M. Oronoz); alicia.perez@ehu.eus (A. Pérez)

0009-0003-7308-5965 (X. Larrayoz); 0000-0003-4248-8182 (A. Casillas); 0000-0001-9097-6047 (M. Oronoz); 0000-0003-2638-9598 (A. Pérez)

2. Materials and methods

In order to address Task 1 of the MentalRiskES competition, data collected from a collection of labeled users along with the messages they had posted over time were utilized[4]. The dataset for training consisted of a total of 485 users, with the following label distributions: 223 users with no detected disorders, 169 users with depression, and 93 users with anxiety, as shown in Figure 1. In total, $17,586$ messages were collected, with an average of 36.26 messages per user and a total of $254,542$ words, averaging 14.47 words per message.

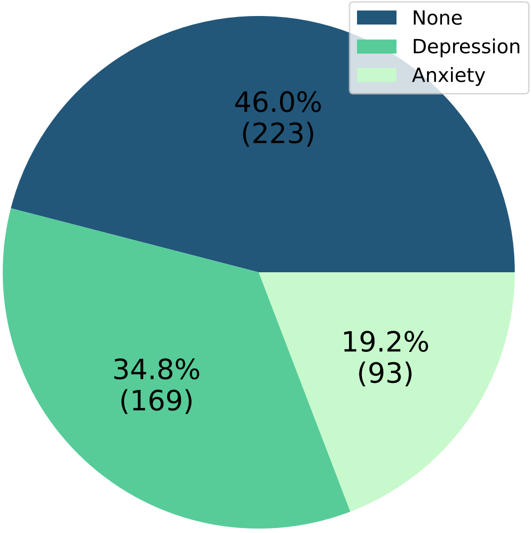

Figure 1: Class distribution of users of the training set

Originally, the class is assigned per user, however, there is no trace of which message or messages motivated the user label. Needless to say, a user classified as Depression or Anxiety might, very well, address in several messages similar topics as other users classified as None. Furthermore, not all messages within positive users show, necessarily, traces of depression or anxiety. However, the aim is to turn to message classifiers in an attempt to detect as early as possible positive users without the need of having processed all the messages. In other words, the approach would require to get messages annotated while we count on user-level classes. With this rationale, we did not find it consistent to label all messages with the corresponding user-label. For all this, we opted for a message-level heuristic re-labeling approach in an attempt to gain consistency at message-level labeling to leverage supervised classification approaches. Our re-labeling approach focused on the cosine similarity of the embeddings vectors of the message. That is, with the similarity of messages m_A and m_B computed as in (1) with $\mathbf{A}$ and $\mathbf{B}$ being the embedding vectors of messages m_A and m_B respectively.

$$Similarity(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

This allowed to identify messages that did not show signs of mental health issues and, accordingly, getting them labeled with the ”none” class. Increased consistency in supervised train has an impact in supervised inference approaches. Relabeling impact is shown by means of Figure 2.

Regarding the supervised learning approach, we involve language models to cope with textual information and a simple neural network to classify the information. In fact, two different language models were evaluated: SBERT (paraphrase-multilingual-mpnet-base-v2) [5] and BETO: Spanish BERT [6]. In an attempt to process the messages, a mean pooling layer was used to obtain a vector representation of each message. Finally, a neural network [7] was applied to perform multi-class classification of the messages into depression, anxiety, or no detected disorder categories. It is important to note that, for a

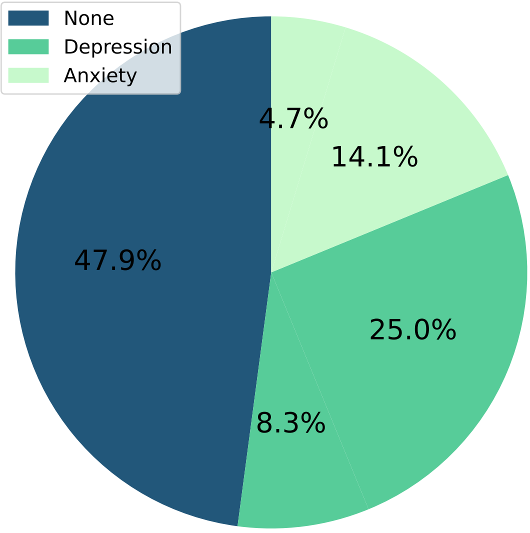

Figure 2: Class distribution of the training set messages. The marked factions are messages that have been reassigned with the None label.

model to assign a specific class to a user, we opted for a accumulated message class-threshold above which the user class was decided.

3. Results

Three different models were implemented to address the task. Each model was configured differently, using different language models and thresholds to make the decision on user classification. Table 1 summarizes the configurations for each run. The first model uses SBERT as the language model and a threshold of 4 for classification. The second model uses BETO as the language model and a threshold of 5. The third model uses SBERT as the language model and a threshold of 3.

Table 1
Submitted Runs denoted as Ixa-Med team: Description of the configurations explored

Run	Encoder	Threshold
0	SBERT	4
1	BETO	5
2	SBERT	3

To evaluate the quality of model predictions, a variety of metrics have been used. These metrics include macro-precision (Macro_P), macro-recall (Macro_R), macro F1-score (Macro_F1) and accuracy. Accuracy, as defined in Equation (2), measures the fraction of correctly predicted instances out of the total number of instances. Macro-precision, calculated according to Equation (3), represents the average precision across all classes, while macro-recall, as defined in Equation (4), computes the average recall across all classes. Finally, the macro F1-score, as described by Equation (5), provides a balanced measure of the model's performance by taking into account both precision and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$Macro\_F1 = \frac{2 \times Macro\_P \times Macro\_R}{Macro\_P + Macro\_R} \tag{5}$$

Our team, denoted as Ixa-Med, ranked seventh among the presented models, as can be seen in Table 2. Despite strong competition, our model achieved an accuracy score of 0.749, demonstrating its competitive performance. Additionally, our model achieved competitive results in several metrics, including a macro-precision of 0.796 and a macro-recall of 0.747. Notably, our team was one of only four teams to surpass a macro F1 score of 0.7, highlighting the difficulty of the task and the effectiveness of our approach in addressing it. However, it is worth mentioning that our score still fell below the baseline proposed by the organization.

Table 2
Results of each team's best model for Task 1

Rank	Team	Run	Accuracy	Macro_P	Macro_R	Macro_F1
1	ELiRF-UPV	2	0.89	0.875	0.88	0.874
3	BaseLine - Roberta Base	2	0.853	0.84	0.843	0.834
5	UnibucAI	0	0.828	0.824	0.808	0.808
8	UNED-GELP	0	0.797	0.792	0.797	0.785
10	Ixa-Med	1	0.79	0.796	0.747	0.749
12	Ixa-Med	2	0.79	0.79	0.733	0.736
13	Ixa-Med	0	0.762	0.763	0.725	0.723
14	BaseLine - Roberta Large	1	0.67	0.786	0.708	0.682
15	UMUTeam	2	0.69	0.701	0.683	0.675
17	BUAP_01	1	0.62	0.692	0.662	0.632
18	BaseLine - mDeberta	0	0.71	0.748	0.645	0.623
19	UC3M-DAD	0	0.578	0.727	0.647	0.601
22	NLP UNED MRES	0	0.557	0.644	0.62	0.561
25	VerbaNex AI	1	0.527	0.598	0.372	0.303
31	Huerta	2	0.47	0.24	0.318	0.231

Table 3
Average carbon emissions of main teams in Task 1

Rank	Team	Duration_mean	Duration_desv	Emissions_mean	Emissions_desv
1	ELiRF-UPV	30.185	15.7	4.79E-04	2.56E-04
2	UnibucAI	7.670	5.491	2.97E-05	2.31E-05
3	UNED-GELP	130.818	91.171	1.84E-04	1.30E-04
4	Ixa-Med	10.238	9.345	8.20E-05	7.58E-05

In Table 3, a comparison of the average emissions produced by the top four teams in the competition is presented. The Ixa-Med team positioned notably well in terms of emissions, achieving an average of 8.20E-05, surpassing most of the participating teams. Specifically, Ixa-Med generated significantly fewer emissions than ELiRF-UPV and UNED-GELP, which presented average emissions of 4.79E-04 and 1.84E-04, respectively. Although UnibucAI had the lowest average emissions, the performance of Ixa-Med is very close and competitive, reflecting an effective balance between efficiency and sustainability.

In summary, although the results obtained are encouraging, there is room for improvement and refinement of our approach in future iterations to remain competitive in the field of online mental disorder detection in the Spanish-speaking context.

4. Conclusions

In this MentalRiskES challenge, various approaches have been presented to address the early detection of mental disorders in Spanish using social media data. The results obtained reflect the growing interest and importance of this research field in identifying and preventing online mental health issues.

Our team, Ixa-Med, has demonstrated its ability to develop a robust and competitive model in the task of mental disorder detection. It is important to note that while we have achieved promising results, there is still room for improvement in our approach. Remaining challenges include optimizing the accuracy and efficiency of the model, as well as exploring new techniques and approaches to address mental disorder detection in a diverse and dynamic online environment.

Besides we found the MentalRiskES competition an invaluable opportunity to advance research in online mental disorder detection in the Spanish-speaking context.

Acknowledgments

This work was elaborated within the framework of LOTU (TED2021-130398B-C22) funded by MCIN/AEI/10.13039/501100011033, European Comission (FEDER), and by the European Union "NextGenerationEU"/PRTR. Besides, this work was partially funded by the Spanish Ministry of Science, Innovation and Universities (EDHIA PID2022-136522OB-C22); by the Basque Government (IXA IT-1570-22) and by EXTEPA within Misiones Euskampus 2.0.

References

[1] World Health Organization, Trastornos mentales, https://www.who.int/es/news-room/fact-sheets/detail/mental-disorders, 2022.

[2] 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.

[3] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. Plaza-del-Arco, M. D. Molina-González, M. T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of MentalriskES at IberLEF 2024: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[4] 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.

[5] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[6] 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.

[7] Bebis, G. and Georgiopoulos, M., Feed-forward neural networks, IEEE Potentials 13 (1994) 27–31. doi:10.1109/45.329294.