Participation of UC3M-DAD on MentalRiskES Task at IberLEF 2024

Dario Muñoz-Muñoz[1], Alvaro Marco-Perez[1] and David Ramirez[1]

[1]*Universidad Carlos III de Madrid*

Abstract

Mental health awareness is increasingly critical, with conditions such as anxiety and depression affecting millions globally. Social media offers a rich data source for early detection of these issues through textual analysis. This study presents the methodology and findings of the UC3M-DAD team's participation in the MentalRiskES task at IberLEF 2024, aimed at identifying signs of anxiety and depression in Telegram chat texts using advanced natural language processing (NLP) and machine learning models. We employed pre-trained language models, including RoBERTuito and BETO Sentiment Analysis models, fine-tuned on a curated dataset of Spanish texts from 465 individuals. Comprehensive preprocessing steps, such as message concatenation and filtering of non-pertinent elements, were implemented to enhance data quality. Results indicate that while the BETO model slightly outperformed RoBERTuito in accuracy, RoBERTuito demonstrated significantly faster processing, making it a more practical choice for real-time applications. The final submission utilized the RoBERTuito model without filtering out neutral texts during preprocessing. The models were evaluated based on accuracy, precision, recall, F1-Score, and early detection metrics such as ERDE5, ERDE30, and latency measures. The participation underscores the potential of NLP and machine learning in mental health monitoring, highlighting both the promise and challenges of accurately detecting mental health conditions from social media text. This work contributes to the development of digital tools for early detection and timely intervention, ultimately aiming to improve mental health outcomes on a global scale.

Keywords

Mental health, anxiety, depression, early detection, social media, Natural Language Processing (NLP), Machine Learning, sentiment analysis, RoBERTuito, BETO

1. Introduction

Mental health awareness has emerged as a critical aspect of public health in recent years. According to the World Health Organization (WHO), mental health disorders such as anxiety and depression affect over 301 million and 280 million people globally, respectively, representing an estimated 4% of the world's population. These conditions not only diminish quality of life but also contribute significantly to the global burden of disease, leading to substantial economic costs and lost productivity. The increasing recognition of the importance of mental health underscores the need for effective strategies to identify and support individuals experiencing these challenges. [1] [2]

In the digital age, social media and other online platforms have become integral to daily communication, providing a rich source of data that can be leveraged to address mental health issues. Recent studies have shown that the language used in social media posts can serve as a window into users' mental states, offering potential for early detection of conditions such as anxiety and depression. For instance, individuals experiencing depression may use more negative language, express feelings of hopelessness, and post less frequently compared to others. [3]

The ability to detect signs of anxiety and depression through textual analysis on social media presents a unique opportunity for early intervention. Early detection is crucial as it allows for timely support and resources, which can significantly improve outcomes for those affected.

[†]These authors contributed equally.

damunozmu@pa.uc3m.es (D. Muñoz-Muñoz); alvaro.marco@alumnos.uc3m.es (A. Marco-Perez); 100405989@alumnos.uc3m.es (D. Ramirez)

https://github.com/dariomnz (D. Muñoz-Muñoz); https://github.com/alvaro-marco (A. Marco-Perez); https://github.com/d4vidram (D. Ramirez)

0009-0009-3574-9189 (D. Muñoz-Muñoz)

Machine learning models and natural language processing (NLP) techniques can analyze large volumes of text data to identify patterns and indicators of mental health issues. By deploying such models, we can proactively reach out to individuals showing signs of distress and connect them with necessary mental health services.

Furthermore, the anonymity and accessibility of social media can encourage individuals to express their feelings more openly than they might in face-to-face interactions, providing a more accurate picture of their mental state. This opens the door for developing digital tools that not only monitor and detect mental health issues but also offer immediate support through chatbots or direct connections to mental health professionals.

Ultimately, leveraging digital tools to analyze social media for signs of anxiety and depression represents a promising approach to enhancing mental health care. By harnessing the power of technology, we can improve early detection, provide timely interventions, and ultimately contribute to better mental health outcomes on a global scale.

In this work, we present the methodology and findings of the UC3M-DAD team's participation in the MentalRiskES task at IberLEF 2024. [4] [5] The primary goal of this task is to identify signs of anxiety and depression within text-based communication extracted from Telegram chats. This involves the application of advanced natural language processing (NLP) techniques and machine learning models to analyze and interpret text data for mental health indicators.

Our approach leverages pre-trained language models tailored for sentiment analysis in Spanish. These models were fine-tuned using a curated dataset of texts from 465 individuals, specifically designed to capture the nuances of language associated with anxiety and depression. Comprehensive preprocessing steps were implemented to enhance data quality, including message concatenation and the removal of non-pertinent elements.

In summary, our participation in the MentalRiskES task at IberLEF 2024 demonstrates the potential of NLP and machine learning in mental health monitoring through social media analysis. This work contributes to the growing field of early detection and intervention for mental health conditions through digital means.

2. Data

In this section, we detail the data architecture and preprocessing steps undertaken for the MentalRiskES task. [6] We describe the data partitioning, preprocessing methods to enhance data quality, and the filtering techniques applied to refine the dataset for model training.

As mentioned beforehand, the dataset comprises Telegram chat texts in Spanish from a large sample of individuals, specifically curated to identify mental health indicators related to anxiety and depression.

It features texts from 465 individuals. Each of these individual's conversation history was collected, encompassing a maximum of 100 texts per person.

Each data instance in the dataset contained the following attributes:

1. **Message ID**: A unique identifier assigned to each message.
2. **User**: The user responsible for the message.
3. **User Label**: Indicates the mental health status of the user, categorized as 'anxiety', 'depression', or 'none'.

The distribution of classes within the dataset is depicted in Figure 1. Notably, the majority of users in the dataset are labeled as 'none', indicating the absence of anxiety or depression indicators.

For model training and evaluation, the dataset was partitioned into training and test sets. The training set comprises 80% of the total individuals (372 people), while the remaining 20% (93 people) constitute the test set. The distribution of texts per individual varies, as illustrated in Figure 2.

Across the dataset, a total of 13,759 texts were allocated for training and 3,552 texts for testing.

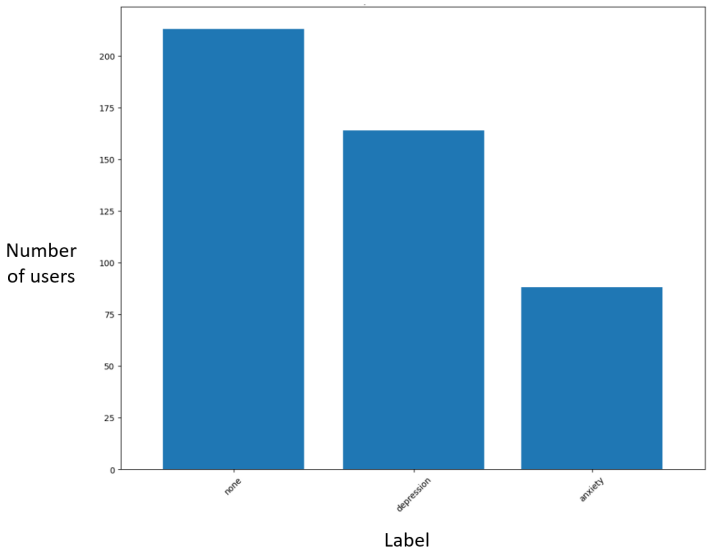

Figure 1: Class distribution of users

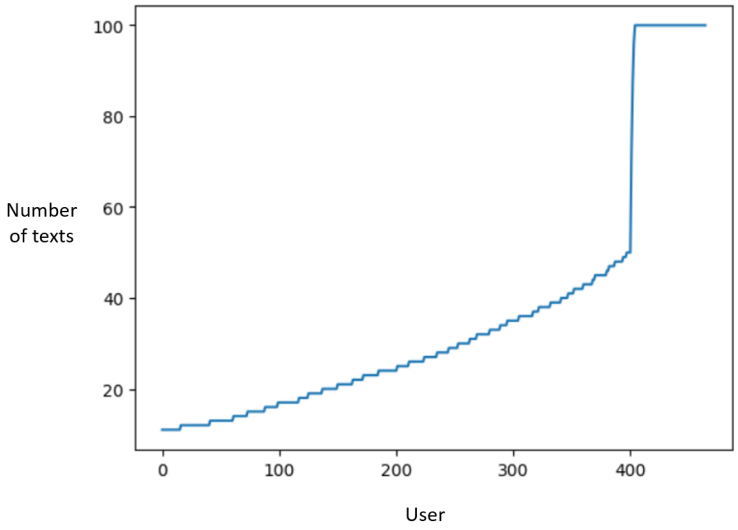

Figure 2: Number of texts per user

3. Additional data

Upon thorough evaluation, it was determined that the supplied data possessed a significant volume to achieve a robust model training. Consequently, no data augmentation techniques or additional datasets were used.

However, a preprocessing procedure was undertaken. This preprocessing involved the concatenation of all the text messages associated with each individual user, thereby creating comprehensive and complete user information.

Furthermore, to mitigate the risk of overloading the model with extraneous information and to enhance computational efficiency, a systematic process was implemented to remove stopwords, symbols, and numerical values. By eliminating these elements, the resulting dataset was refined to contain only pertinent textual information essential for model training. This approach enhanced the model's ability to extract meaningful patterns and insights from the data.

Given the substantial volume of texts contributed by certain users, which may surpass the token limit for training, we explored the concept of filtering out potentially irrelevant texts—those lacking

distinctiveness in identifying signs of anxiety or depression. In certain experiments, we employed a zero-shot classifier to exclude non-mental health-related texts from the final concatenation. The considered categories were 'emociones' and 'neutral'. A text which is supposed to be related to emotions should get a greater probability of fitting the category 'emociones', and the opposite way for a neutral or irrelevant text. In some experiments described below, texts with a high probability of being considered neutral are filtered out, varying the level of restrictiveness, and not included in the concatenation.

4. Pre-trained models

In our study, we have utilized two pre-trained language models, each tailored to analyze sentiments in Spanish text. The first model, RoBERTuito Sentiment Analysis, available at Hugging Face [7], was trained using the TASS 2020 corpus, consisting of approximately 5,000 tweets across various Spanish dialects. This model is based on RoBERTuito, a variant of RoBERTa designed specifically for Spanish tweets. One notable feature of this model is its input size of 128 tokens, which allows for faster processing compared to models with larger input sizes.

The second model, BETO Sentiment Analysis, also accessible on Hugging Face [8], was trained on the same TASS 2020 corpus. However, it is built upon BETO, a variant of BERT tailored for the Spanish language. Unlike RoBERTuito, this model accommodates larger input sizes, allowing up to 512 tokens. While this grants it the capacity to capture more context, it operates slightly slower than RoBERTuito due to the increased input size.

These models were chosen for their pre-training on sentiment-related tasks, a domain closely intertwined with mental health. Sentiment analysis serves as a valuable tool in understanding the emotional states expressed in text, making these models particularly relevant for our investigation into mental health-related topics.

There was also another pre-trained model that was used in some experiments to filter texts that might be considered irrelevant for the tasks, as described in Section 3. This one is the BART large mNLI available at Hugging Face. It is a NLI-based Zero Shot Text Classification model that uses data from Facebook, and is able to associate a text to class names in two languages: English and Spanish.

5. Experiments conducted and training parameters

The hyperparameters used for fine-tuning the model were the following:

- Learning rate: $2e^{-5}$
- Train batch size: 20
- Eval batch size: 20
- Epochs: 2
- Weight decay: 0.01

We performed five different experiments for Task 1 by trying different configurations with the pre-trained models mentioned in Section 4.

For the Experiment 1, a fine-tuning of the RoBERTTuito Sentiment Analysis model was performed, with the available data and the hyperparameters mentioned above.

For the Experiment 2, a similar procedure to Experiment 1 was followed, but using the BETO Sentiment Analysis model instead. Experiments 3, 4 and 5 are variations of Experiment 2, including an additional step on the pre-processing stage by filtering out irrelevant texts. As explained in Section 3, this is carried out by labeling every text according to a Zero Shot Classifier, considering two possible categories: 'emocional' or 'neutral'. For these experiments, different degrees of restrictiveness was chosen.

On Experiment 3, only the texts with more than a 65% chance of being related to emotions are concatenated, significantly reducing the tokens size, but taking the risk of excluding important information,

as the classifier is far from perfect and might label relevant information as neutral. On Experiment 4, the texts with more than a 50% chance of being relevant are chosen. And on Experiment 5, only the texts with more than a 60% chance of being neutral were excluded on the training, not reducing the data size in a big way, but also not taking a big risk on losing important information.

In every experiment, we decided to remove stop words. Also, for every experiment, we chose that the tokenizer performed left truncation, aiming to gradually remove the oldest messages from the history. The reason behind this decision is that we believe that individuals experiencing either anxiety or depression will exhibit certain patterns within a specific number of messages. Table 1 shows the accuracy obtained with the available test data for each described experiment.

Experiment	Accuracy (%)
1	0.68
2	0.71
3	0.66
4	0.71
5	0.69

Table 1
Comparison of experiments for Task 1

As the results show, Experiments 3-5 did not perform better than Experiment 2. This illustrates that filtering out neutral texts did not significantly improve the model's performance. The accuracy remained fairly consistent across experiments, suggesting that the inclusion of all available texts, with thorough preprocessing, provided the most reliable results.

So, for the Task 2, this idea was discarded, and only two experiments were carried out. Experiment 1 used the RoBERTTuito Sentiment Analysis model, and Experiment 2 used the BETO Sentiment Analysis model. The results for these experiments are shown on Table 2.

Experiment	Accuracy (%)
1	0.80
2	0.81

Table 2
Comparison of experiments for Task 2

The achieved accuracy indicates that the BETO Sentiment Analysis model slightly outperformed the RoBERTTuito Sentiment Analysis model. However, the latter exhibited significantly faster performance and yielded results practically on par with the former. Thus, it emerges as the more viable choice. Consequently, for the submission, the selected model was the RoBERTTuito Sentiment Analysis one, without filtering out neutral texts during pre-processing.

6. Analysis of results

The final results of the performance of our model in the competition are discussed on this section.

As presented in [9], the two tasks are evaluated according to two different criteria: absolute classification and early detection efectiveness. Table 3 shows the final results according to the following metrics, which measure the performance in terms of absolute classification: accuracy, Macro Precision (Macro-P), Macro Recall (Macro-R), and Macro F1-Score (Macro-F1). Table 4 shows the results in terms of early detection efectiveness, considering the computed metrics: Early Risk Detection Error @ 5 (ERDE5), Early Risk Detection Error @ 30 (ERDE30), Latency of True Positives (latencyTP), speed, and Latency-weighted F1.

As we decided to only submit the model that performed the best in our tests, there is no distinction between the three different runs that we were allowed to do, and so only a result per task is discussed. That result occupies three consecutive positions in the ranking, as it is reflected on the tables.

Task	Position	Accuracy	Macro-P	Macro-R	Macro-F1
1	19-21 out of 33	0.578	0.727	0.647	0.601
2	13-15 out of 19				

Table 3
Results - absolute classification metrics

Task	Position	ERDE5	ERDE30	latencyTP	speed	latency-weightedF1
1	18-20 out of 33	0.205	0.133	1	1	0.811
2	14-16 out of 19					

Table 4
Results - early detection metrics

The results indicate that our models were effective in identifying signs of anxiety and depression in text-based communication. However, the relatively moderate performance suggests there is room for improvement, particularly in refining the preprocessing steps and exploring additional features that may better capture the nuances of mental health indicators in textual data.

Overall, our participation in the MentalRiskES task highlights the potential of NLP and machine learning in mental health monitoring, but also underscores the challenges involved in accurately detecting mental health conditions from social media text.

7. Error analysis

Our participation in the MentalRiskES task provided valuable insights into the strengths and limitations of our approach. This section presents a detailed analysis of the failed predictions, potential improvements, and lessons learned.

7.1. Failed Predictions and Their Characterization

One of the primary challenges we faced was the inherent ambiguity in natural language. Many failed predictions occurred in cases where the language used in the texts was ambiguous or subtle, making it difficult for the models to accurately identify signs of anxiety or depression. For example, users might use humor or sarcasm to mask their true feelings, which the models often misinterpreted.

Another issue was our models sometimes failing to capture the full context of a conversation. Mental health indicators often depend on nuanced understanding of ongoing discussions. This limitation was particularly evident in cases where the models misclassified messages that required a broader conversational context to interpret correctly.

The dataset had a significant imbalance, with the majority of users labeled as 'none' for mental health indicators. This imbalance likely caused the models to become biased towards predicting the absence of mental health issues, leading to false negatives. Users exhibiting subtle or less frequent signs of anxiety or depression were often overlooked.

While our preprocessing steps aimed to enhance data quality, they might have inadvertently removed important contextual information. For instance, the removal of stopwords and symbols, although intended to streamline the text, may have led to the loss of nuances crucial for accurate mental health detection.

7.2. Possible Improvements

To address the issue of missed context, future work could involve developing models capable of analyzing longer text sequences or entire conversations. This would allow the models to better understand the context in which certain statements are made, improving prediction accuracy.

Implementing data augmentation techniques to generate more examples of anxiety and depression indicators could help mitigate the imbalance in the dataset. Techniques such as synthetic data generation

or oversampling could provide the models with more balanced training data, enhancing their ability to detect less common mental health indicators.

Integrating advanced sentiment analysis tools can help capture the emotional tone of the messages more accurately. This could involve using models specifically trained to understand nuanced sentiments and emotions, which are often key indicators of mental health status.

Refining our preprocessing approach to preserve essential contextual elements while still removing irrelevant information could strike a better balance. This might include selectively removing stopwords or symbols based on their contextual relevance rather than applying a blanket removal approach.

7.3. Lessons Learned

One of the critical takeaways from our analysis is the importance of context in mental health detection. Future models should prioritize maintaining and understanding conversational context to improve prediction reliability.

Using a diverse set of models, each with strengths in different areas of text analysis, can help cover a broader range of linguistic nuances. Ensemble methods that combine predictions from multiple models might provide a more comprehensive analysis.

The field of NLP and mental health detection is rapidly evolving. Staying updated with the latest advancements and incorporating new techniques and models can significantly enhance our approach. Continuous learning and adaptation are essential for developing effective mental health detection tools.

In conclusion, our error analysis highlights the challenges and areas for improvement in using NLP models for mental health detection. By addressing these issues, we can develop more robust and accurate systems that contribute meaningfully to early detection and intervention efforts in mental health care.

8. Conclusions

Our participation in the MentalRiskES task at IberLEF 2024 has been an enriching and educational experience. Throughout this journey, we gained significant insights into the implementation and fine-tuning of NLP models specifically designed for detecting mental health indicators. The challenges we encountered and overcame highlighted the complexity and importance of developing accurate and reliable algorithms in the field of mental health.

This endeavor underscored the critical role that advanced NLP techniques and machine learning models play in early detection and intervention for mental health conditions. By analyzing text-based communication from social media platforms, we can potentially identify signs of anxiety and depression, facilitating timely support and resources for affected individuals.

Our models demonstrated promising results, but there is still considerable room for improvement. The lessons learned from this task, including the importance of data preprocessing, model selection, and parameter tuning, will guide our future research and development efforts. We are committed to refining our approaches to enhance the accuracy and reliability of mental health detection tools.

Participating in the MentalRiskES task provided us with valuable experience and deeper understanding of the nuances involved in this critical area of study. We thoroughly enjoyed the collaborative and competitive aspects of this challenge and look forward to continuing our contributions to advancing mental health detection through innovative NLP solutions.

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles https://github.com/borisveytsman/acmart and to the developers of Elsevier updated LaTeX templates https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates.

References

[1] World Health Organization, Anxiety disorders, 2023. URL: https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders, accessed: 2024-05-17.

[2] World Health Organization, Depressive disorder (depressión), 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression, accessed: 2024-05-17.

[3] N. Akhther, P. Sopory, Seeking and sharing mental health information on social media during covid-19: Role of depression and anxiety, peer support, and health benefits, Journal of Technology in Behavioral Science 7 (2022) 211–226. doi:10.1007/s41347-021-00239-x.

[4] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[5] 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).

[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, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://huggingface.co/pysentimiento/robertuito-sentiment-analysis.

[8] 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 (2021). URL: https://huggingface.co/finiteautomata/beto-sentiment-analysis.

[9] 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, Official results of mentalriskes at iberlef 2024: Early detection of mental disorders risk in spanish, 2024.