Mental Disorders Detection with Immediate Message using RoBERTa

Chenghao Hu, Xiaobing Zhou*

School of Information Science and Engineering, Yunnan University, Kunming 650500, Yunnan, China
** Corresponding author: zhouxb@ynu.edu.cn*

Abstract

Nowadays, all kinds of instant message applications have become integral parts of our daily lives. With the quickening pace of society and the increasing work pressure, more and more people suffer from mental disorders such as eating disorders, depression, and unknown disorders. In MentalRiskES 2023, participants tried to detect mental disorder risk early in Spanish, where the corpora are attained from the chat message records of Telegram. This paper describes the participation of the group GetitDone on the 2a subtask. Our team uses BETO, also called Spanish BERT, pretrained on a large Spanish corpus, as our base model. We put efforts into preprocessing the given data and making changes to the classification part of the model.

Keywords 1

depression detection, disorder detection, dataset preprocessing, BERT, RoBERTa

1. Introduction

Over the past decade, instant messaging apps have gradually become integral to people's lives. Larger and larger message streams contain more information than it seems. Though multimedia message has been a significant form of expression, which can also be used in multi-model sentiment analysis, the text is still the most important carrier of human language and is the primary way we express ourselves. For example, when we send a message to our friend, it may contain some subconscious thoughts you will never notice. The major goal of this task is to capture potential mental disorders through the chats between users, which means early detection. It can be noticed that some relevant evaluation campaigns [1] have been held previously, but almost all of them are for English. Hardware limitation, unfamiliarity with Spanish, and lack of corpus are challenges for our group.

The importance and urgency of mental health have become increasingly recognized in today's society. Mental health issues, such as depression, profoundly impact individuals. The prevalence of mental health disorders is alarmingly high, with millions worldwide affected by these conditions. The COVID-19 pandemic has further highlighted the significance of mental health [2]. Early detection and intervention are crucial in effectively addressing mental health issues, especially for those who have already developed mental health problems but are unaware of them. Identifying signs and symptoms early on allows for timely support, treatment, and prevention of further deterioration. It can help individuals regain control over their lives.

It is worth noting that in addition to the prediction results to be submitted, it is also required to submit information about CO2 emissions, which puts further requirements on the performance of the algorithm and model. Our group works on the binary classification task of depression detection, which is subtask 2a. Unlike conventional sentiment analysis, whose training and testing data are just some sentences or paragraphs, the data provided here is a set of chat messages, which can be regarded as a series of related

IberLEF 2023, September 2023, Jaén, Spain
EMAIL: dhxxhch@gmail.com (H. Chenghao)
ORCID: 0009-0000-0929-8024 (H. Chenghao)


sentences. This paper uses some techniques to preprocess the training data and organize the data before training.

In this paper, we first mention some related work on depression detection and carbon emissions of AI models and then introduce the dataset used in the pre-trained model and some characteristics of the given dataset. In this section, the preprocessing method this paper used is also introduced. The model section describes the model we proposed in detail, including the structure of the model and some hyperparameters used in the training. Lastly, we analyze the results from three aspects, draw a conclusion about our work, and put forward a few directions for future work.

2. Related Work

The detection of depression in online user-generated content has gained significant attention in recent years. Various studies have explored different approaches and techniques to address this challenge. Early research focused on using linguistic features and sentiment analysis to identify depressive symptoms in text data. Nadeem et al. [3] employed machine learning algorithms to classify depression-related posts on online forums based on lexical and syntactic patterns. Similarly, Gautam, and Yadav [4] utilized sentiment analysis to detect depressive language patterns in Twitter data.

With the advancements in natural language processing (NLP) and deep learning, researchers started leveraging pretrained language models for depression detection. For instance, Bucur et al. [5] applied a fine-tuned BERT model to classify depression-related text obtained on Reddit. Additionally, Wolk et al. [6] explored the use of GPT for identifying depression symptoms in social media posts.

Another research interest involves analyzing social network structures and behavioral patterns. Islam et al. [7] investigated the impact of social network characteristics on depression detection, highlighting the importance of social connections in predicting mental health conditions.

Furthermore, researchers have explored integrating multimodal data, such as text, images, and audio, to enhance depression detection. Jahan et al. [8] proposed a multimodal approach that combined images, videos, and text to improve the accuracy of depression detection in Twitter and Facebook posts.

Overall, the existing literature showcases a wide range of methodologies, including linguistic analysis, deep learning models, social network analysis, and multimodal fusion, for detecting depression in online user-generated content.

When we searched for relevant papers, we found that fewer studies have been conducted on depression detection in Spanish compared to more widely spoken languages like English, which make the MentalRiskES evaluation campaigns more meaningful. This paper aims to propose a fast and resource-efficient method to complete depression detection.

Carbon emissions associated with AI have become a topic of concern in recent years. The rapid growth of AI applications, especially deep learning models, requires significant computational resources that contribute to increased energy consumption and, in turn, carbon emissions. A study by Strubell et al. [9] analyzed the carbon footprint of training large language models. ECO2AI [10] can also be used to track the carbon emissions of machine learning models.

3. Dataset

The provided training dataset is separated by subject. Every subject file contains the conversation text between the target subject, whom we'd like to predict whether suffers from depression or not, and other people. Each message in the conversation is also attached with an ID number and a datetime, which makes every sentence identical and chronological.

As with conventional text cleaning, we especially take care of hashtags, special symbols (e.g., currency symbols), URLs, emoji, and repetitive characters first.

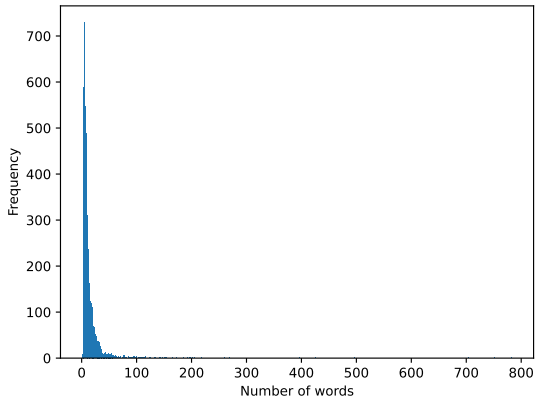

Figure 1: Histogram of post lengths

When checking the training dataset, we found some of the messages are so tricky because of their large scale in length. Figure 1 shows the frequency of sentences in terms of the number of words in the sentence. The maximum length of the message sentence is 4279, which has about 750 words. Figure 2 shows the method we used to handle this situation.

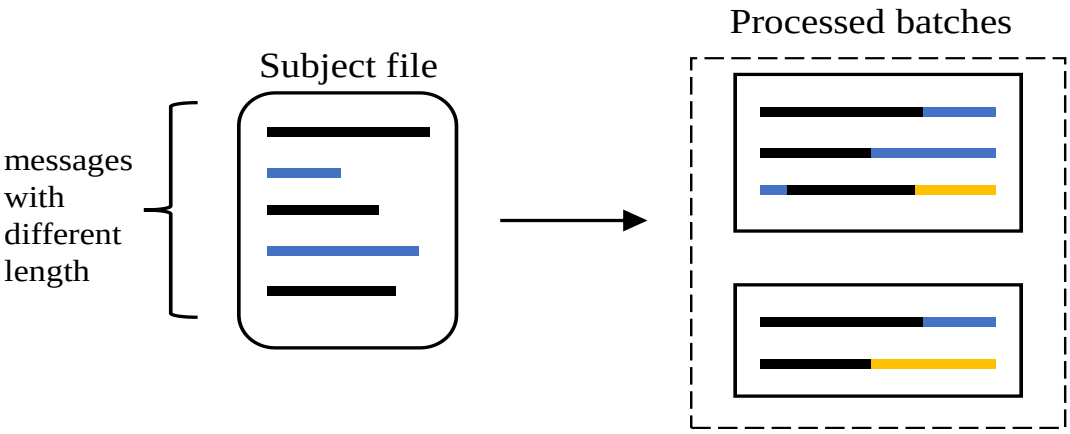

Figure 2: The method to preprocess the message with different length

Generally speaking, we clip and join the message into the same length first, then pad the last line with padding tokens. Additionally, in the process of clipping and joining, we pick up a random percentage of messages at the beginning of every subject file. For example, in Figure 2, there are five messages in a certain subject file. We pick up all of the messages to produce the first batch and 60% to produce the second one.

The reason we clip and join these messages is as follows: We believe that the information about the depressions does not distribute evenly across messages, so we try to help the model to focus more on the information that actually reflects whether or not the subject is depressed. The clipping and joining operation could relatively increase the density of this kind of information.

In the testing phase, we build up and append a subject file for every message we received. And when making a prediction on a certain subject, we will put the messages received in the current round together with the ones received previously and then feed them into the model.

4. Model

Our model is based on the model in Python library pysentimiento [11], specifically, the pretrained model robertuito-sentiment-analysis, which was trained based on a model called RoBERTuito [12]. This model is based on RoBERTa [13] and trained on a Spanish tweets corpus TASS 2020 [14].

The architecture of the model we proposed is shown in Figure 3. For the tokenizer part, we use the default setting in the pretrained model robertuito-sentiment-analysis, which distinguishes 30002 kinds (including the special tokens such as [PAD], [SEP], etc.) of word index. The encoder is the primary improved part of our proposed model architecture. It begins with a 12-head self-attention, followed by a GRU, and finally, a fully connected layer. As for the classifier, we use a fully connected layer, a dropout layer and a linear layer.

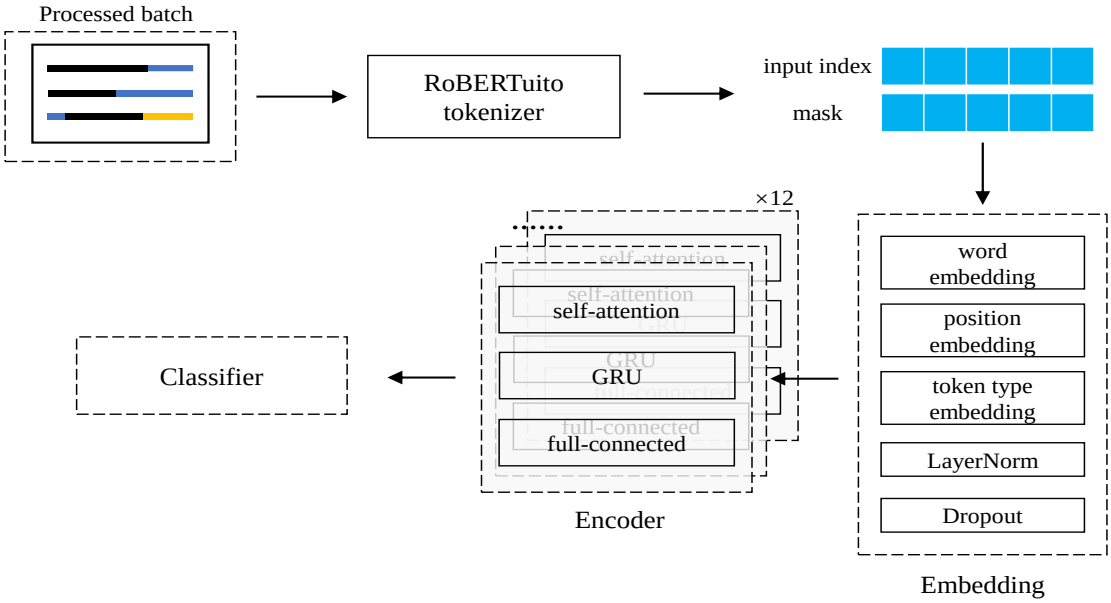

Figure 3: Proposed model architecture

In the classification part, we use a three-class classifier, the same as conventional sentiment analysis has done. For this task, the sum of neutral and positive probabilities is considered the non-depressed probability, and the negative probability is considered the depressed probability. When we process a new message using the method mentioned above, if the depressed probability exceeds the non-depressed probability, which means the depressed probability is greater than 50 percent, we report the depression immediately.

Table 1
The details of encoder

Name	Layers included	Size/Probability
attention	query (Linear)	768, 768
	key (Linear)	768, 768
	value (Linear)	768, 768
attention-out	dense (Linear)	768, 768
	LayerNorm	768
	Dropout	0.1
recurrent	GRU	768, 768
	Dropout	0.1
recurrent-out	dense (Linear)	768, 768
	LayerNorm	768
	Dropout	0.1
intermediate	dense (Linear)	768, 3072
	GELUActivation	
intermediate-out	dense (Linear)	3072, 768
	LayerNorm	768
	Dropout	0.1

The following is a detailed description of the model's details. After getting the tokens of every sentence, we will put the input ids and attention mask into the embedding part. In the embedding part, we use word embedding, position embedding, and type embedding. All of the embedding dims are taken to be 768, while the size of the dictionary is set to 30002, 130, and 1, respectively, which means the max length of a sequence is 128. Next is an LN layer, whose eps is set to 10---12, and a dropout layer, whose probability is set to 0.1.

As for the encoder part, every self-attention part is composed of attention, recurrent and intermediate. Table 1 shows the details of the encoder. Please check out the code for our model on GitHub [15] for more information.

In the fine-tuned training phase, we use a common set of configurations combined with our actual hardware conditions, including a batch size of two, an optimizer of AdamW, and a learning rate of 0.00002. It is hard to determine the number of epochs during the training. We first try to train for twenty epochs, but the accuracy starts to fall after about fifteen epochs, so finally twelve epochs are believed to be best for our model.

5. Results

The tasks in MentalRiskES are all about the online problem of mental disorder detection, so our group focuses more on the speed and latency of the detection. According to the official results [16], in task 2a, we get an excellent result, ranking first, on latencyTP and speed, which is shown in Table 2. For the application scenario of this task, we believe that early detection is even more important than accuracy because people should learn whether they have depression or not as soon as possible. The results have proved that our strategy of judging time works on the speed of detection.

Table 2
Latency-based evaluation

Team	LatencyTP	Speed	latency-weightedF1
GetitDone	2.000	0.984	0.627

In terms of accuracy, the accuracy is greater than 60 percent, ranking twentieth, which can be regarded as an acceptable result. Considering the practical application scenario, stability is also very important other than speed and latency. The Macro-average indexes, ranking 18th in terms of Macro-F1, are almost the same in the evaluation, which indicates that our method used for this task is barely stable.

Table 3
Classification-based evaluation

Team	Accuracy	Macro-P	Macro-R	**Macro-F1**
GetitDone	0.611	0.628	0.622	0.609

In the training phase, we get an accuracy of 0.62 on the training dataset, which is very close to the result in the evaluation phase shown in Table 3. This phenomenon shows that the bottleneck is in the dataset or the model itself. However, as we can see in the ranking, the highest accuracy is 0.738, which is far below other sentiment analysis tasks. This result may indicate that the dataset is not large enough for the model to learn the characteristics of the depression feature. Besides, overfitting is also present in our model. It will happen after training for about 12 epochs, which is also an area for further improvement.

Table 4
Some subjects with wrong prediction

Subject	Golden truth	Prediction
subject9	0	1
subject34	1	0
subject199	1	0
.....		
subject205	1	0
subject220	1	0

Our team check out all of the subjects whose predictions were wrong, which is shown in Table 4, and find that almost every one of them has raw Unicode code starting with "\u". But intuitively, the impact of a few words in the middle of a message on the overall judgment should be very small. This also shows that the model proposed in this paper relies very strongly on the continuity of word senses in sentences.

Additionally, in our assumption, we should try our best effort to detect the subject who is suffering depression, but as a result, the wrong predictions are almost evenly split, which is also a direction that can be improved.

Table 5
Carbon footprint got from codecarbon library

	Emissions	Emissions rate	CPU power	GPU power	RAM power
mean	1.99×10^{-6}	1.15×10^{-7}	22.50	6.03	0.26
min	1.70×10^{-7}	1.76×10^{-8}	22.50	1.00	0.11
max	5.89×10^{-5}	3.50×10^{-6}	22.50	10.91	0.60

Table 5 shows the mean, minimum, and maximum carbon footprint statistics of the predictions. Actually, the statistical data from the CodeCarbon library is really rough, which may not reflect the actual situation. During the predictions, our model only uses the CPU for the low carbon and generality considerations.

6. Conclusions and Future Work

Generally speaking, this paper represents the GetitDone participation for the MentalRiskES 2a task. We got inspiration from the RoBERTa model and used a model pretrained on Spanish corpora. With the hardware limitations, it's difficult to do more pretraining on the extra depression corpus in Spanish. During the experiment, we tried some common classifiers, but no much better results were obtained. So we use a linear classifier in our model.

In future work, three aspects can be considered to improve the experiment as follows.

1. Collect more depression detection corpora to train the existing model.
2. Improve the classifier to make a more reasonable decision.
3. Work out a structure to hold the data that can reflect more data features.

7. References

[1] Martın-Rodilla, P., Losada, D. E., & Crestani, F. (2022, August). Overview of eRisk 2022: Early Risk Prediction on the Internet. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings (Vol. 13390, p. 233). Springer Nature.

[2] Latoo, J., Haddad, P. M., Mistry, M., Wadoo, O., Islam, S. M. S., Jan, F., ... & Alabdulla, M. (2021). The COVID-19 pandemic: an opportunity to make mental health a higher public health priority. BJPsych open, 7(5), e172.

[3] Nadeem, M., Horn, M., Coppersmith, G., & Sen, S. Identifying Depression on Twitter.
[4] Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh international conference on contemporary computing (IC3) (pp. 437-442). IEEE.
[5] Bucur, A. M., Cosma, A., & Dinu, L. P. (2021). Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT, CEUR-WS.org, online http://ceur-ws.org/Vol-2936/paper-77.pdf
[6] Wołk, A., Chlasta, K., & Holas, P. (2021). Hybrid approach to detecting symptoms of depression in social media entries. Retrieved from https://aisel.aisnet.org/pacis2021/192
[7] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. Health information science and systems, 6, 1-12.
[8] Jahan, R., & Tripathi, M. M. (2022). Multimodal depression detection using machine learning. In Artificial Intelligence, Machine Learning, and Mental Health in Pandemics (pp. 53-72). Academic Press.
[9] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics.
[10] Budennyy, S., Lazarev, V., Zakharenko, N., Korovin, A., Plosskaya, O., Dimitrov, D., Arkhipkin, V., Oseledets, I., Barsola, I., Egorov, I., Kosterina, A., & Zhukov, L. (2022). Eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI. arXiv e-prints, arXiv:2208.00406.
[11] Perez, J., Giudici, J., & Luque, F. (2021). pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. arXiv e-prints, arXiv:2106.09462.
[12] Perez, J., Furman, D., Alonso Alemany, L., & Luque, F. (2022). RoBERTuito: a pre-trained language model for social media text in Spanish. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 7235–7243). European Language Resources Association.
[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov. (2020). RoBERTa: A Robustly Optimized BERT Pretraining Approach.
[14] García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M. A., Del Arco, F. M. P., Montejo-Ráez, A., Jiménez-Zafra, S. M., ... & Chiruzzo, L. (2020, September). Overview of TASS 2020: Introducing emotion detection. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain (pp. 163-170).
[15] Hu, G. 2023MentalRiskES_task2a_model. https://github.com/GeorgeHu6/OpenCode/tree/main/2023MentalRiskES
[16] Mármol-Romero, A., Moreno-Muñoz, A., Plaza-del-Arco, F., Molina-González, M., Martín-Valdivia, M., Ureña-López, L., & Montejo-Ráez, A. (2023). Overview of MentalRiskES at IberLEF 2023: Early Detection of Mental Disorders Risk in Spanish. Procesamiento del Lenguaje Natural, 71.