I2C-UHU at MentalRiskES 2023: Detecting and Identifying Mental Disorder Risks in Social Media using Transformer-Based Models

Laura Vázquez Ramos, Carlos Moreno García, Jacinto Mata Vázquez and Victoria Pachón Álvarez

I2C Research Group. University of Huelva, Spain

Abstract

This paper presents the approaches proposed by the I2C Group to address *MentalRiskES: Early Detection of Mental Disorder Risks in Spanish*, as part of IberLEF 2023. Our proposal involves developing distinct transformer-based classifiers to tackle three specific tasks: i) Task1a: Binary classification for the detection of eating disorders, ii) Task1b: Simple regression for the detection of eating disorders, and iii) Task2c: Multiclass classification for the detection of depression. The main approach consisted of fine-tuning pre-trained transformer-based models. For the binary tasks, diverse methodologies were employed to predict users based on the predictions obtained from their individual messages. For the multiclass task, data augmentation approaches were used to balance the minority classes messages. The final submitted predictions achieved a Macro-F1 score of 0.641 for Task1a, ranking 19th out of 22 participants; an RMSE of 0.24 for Task1b, ranking 4th out of 17 participants; and a Macro-F1 score of 0.232 for Task2c, ranking 4th out of 10 participants.

Keywords

Mental Health, Early Detection, Natural Language Processing, Deep Learning, Mental Disorders, Transformer-based Models

1. Introduction

According to a recent report by the World Health Organization, 1 in every 8 people in the world suffers from a mental disorder. The COVID-19 pandemic has raised the prevalence of anxiety and depression to over 26% in just one year. Suicide ranks as the fourth leading cause of death among 15–29-year-olds. The organization considers early identification as a key and effective intervention to prevent these problems.

As a result, there is a growing interest in detecting and identifying mental disorders in social media streams [1]. This addresses the demand from society due to the significant increase in these issues among the population, including various types of mental risks such as eating disorders, dysthymia, anxiety, depression, suicidal ideation, and others.

This paper presents our approach and systems description for *MentalRiskES: Early Detection of Mental Disorder Risks in Spanish*, developed for IberLEF 2023 to identify mental disorder risks at an early stage.

Given the success and popularity of transformers [2], all the developed models in this study are based on this technology. To prepare the training data, each user message was individually labelled. For the final prediction of the user, a percentage method was applied to obtain the final prediction.

IberLEF 2023, September 2023, Jaén, Spain

laura.vazquez005@alu.uhu.es (L. Vázquez); carlos.moreno831@alu.uhu.es (C. Moreno); mata@uhu.es (J. Mata); vpachon@dti.uhu.es (V. Pachón)

Additionally, data augmentation [3] was used for multiclass classification. Moreover, a study of hyperparameters was conducted to ensure efficient training of the models.

The subsequent sections of this paper are structured as follows: Section 2 provides an exploration of related works, examining previous research carried out in the respective field. Section 3 outlines the dataset and tasks. Section 4 details the different approaches used to achieve each task. Section 5 presents the results and analysis obtained from the conducted experiments. Lastly, Section 6 offers conclusions derived from the study.

2. Related works

The significant increase in the number of mental disorder cases due to the COVID-19 pandemic has attracted a growing number of researchers to this field of study, resulting in various proposals and methodologies.

These proposals and methodologies are closely intertwined with the exponential growth of social media worldwide. The information derived from this vast amount of social media data is crucial for building these systems and conducting experiments.

A survey was conducted to examine case studies involving experiments [4], encompassing a total of forty approaches, all of which are related to the detection of mental disorders. This exemplifies the current global importance of addressing this health issue. Furthermore, there have been studies specifically focused on advancing the prediction and detection of eating disorders within these case studies [5, 6].

3. Datasets and tasks

The dataset supplied by the organizers comprises a collection of labelled messages transmitted to group channels on the widely used Telegram platform. The corpus was split into three distinct subsets, each associated with a distinct disorder. Each dataset encompasses several hundreds of users, with an average message count of 50 per user.

- Corpus eating disorders (Task 1): 175 users for training, and 150 for testing.
- Corpus depression (Task 2): 175 users for training, and 150 for testing.

For each user, a json file was provided. Additionally, a distinct file containing the labels of each user was provided as well. By utilizing both files, a training dataset was created. The labelling scheme applied to each user's messages was consistent with the user's label. For instance, if a user was assigned label 1, all their messages were individually labelled as 1.

Two distinct training datasets were created, one for Task 1 and another for Task 2. The dataset for Task 1 consists of five columns: *user*, *message*, *date*, *label_1a* (label for Task 1a), and *label_1b* (label for Task 1b). Table 1 shows the structure of the training dataset for Task 1.

Table 1

Example of training dataset for Task 1

User	message	date	label_1a	label_1b
subject165	David pero lo que veo difícil es eso. Recortar de algo sin q me afecte los otros valores	2019-02-27 15:06:23	1	0.5
subject182	Liz, el tuyo no ha hecho aún un año	2019-03-03 16:15:16	0	0.1
subject222	No te enfades. Creo que todos opinamos como lo haces tú	2019-08-05 09:35:10	0	0.3

On the other hand, the dataset for Task 2 is composed of 4 columns as it can be seen in Table 2: *user*, *message*, *date*, *label_2c* (label for Task 2c).

Table 2
Example of training dataset for Task 2

user	message	date	label_1c
subject305	Yo tengo depresión crónica con trastorno de ansiedad	2021-02-27 19:35:37	control
subject257	Y a ti te ocurre algo además de cantar muy bien	2020-09-10 00:06:21	suffer+ag ainst
subject101	ok no puedo dormir me siento horrible	2021-02-09 10:58:47	suffer+in favour

After constructing the datasets, they were split into three sets: train, test, and valid. The train and valid sets were allocated 80% of the data, while the test set received the remaining 20%. Within the train dataset, 10% was further allocated for validation. Table 3 shows the distribution of examples for each dataset.

Table 3
Distribution datasets for Task 1 and Task 2

Task	Train dataset		Valid dataset		Test dataset	
	Users	Messages	Users	Messages	Users	Messages
Task 1	126	4398	14	522	35	1011
Task 2	126	4450	14	654	35	1144

The distribution of the classes in each created dataset can be observed in Table 4 and Table 5. Furthermore, Table 6 provides an overview of the distribution specifically for Task 2c.

Table 4
Class distribution of datasets in Task 1a.

Class	Train dataset	Valid dataset	Test dataset
0	2575	283	541
1	1823	239	470

Table 5
Class distribution of datasets in Task 1b.

Class	Train dataset	Valid dataset	Test dataset
[0 - 0.3]	2440	233	441
[0.4 - 0.6]	444	150	162
[0.7 - 1]	1514	139	408

Table 6
Class distribution of datasets in Task 2c.

Class	Train dataset	Valid dataset	Test dataset
Control	2275	237	623
Suffer+in favour	1062	210	267
Suffer+against	1016	195	231
Suffer+other	97	12	23

As it was described before, this paper is focused on Task 1 (Task 1a and Task 1b) and Task 2 (Task 2c). Task 1a is a binary classification in which it must be detected if the user suffers from anorexia or bulimia. Labels will be 0 for “control” (negative, the user does not suffer from eating disorder) or 1 for “suffer” (positive). Task 1b is a simple regression in which a probability for the user to suffer anorexia or bulimia must be predict. A value of 0 means 100% negative and a value of 1 would be 100% positive. Finally, Task 2c is a multiclass classification. The system must predict one of the four classes (“suffer+against”, “suffer+in favour”, “suffer+other”, “control”).

4. Methodology and experiments

In this section, we provide an in-depth explanation of the various methodologies employed to address the tasks. We will delve into the specific experiments conducted for Task 1 and Task 2, outlining the steps taken and the techniques employed. Additionally, we present the results obtained from these experiments, which will be thoroughly analyzed and discussed in the subsequent sections. In this section, we aim to provide a clear understanding of the experimental framework employed in our study.

4.1. Task 1: Eating disorders detection

To perform Task 1a and Task 1b, different transformer-based models, available in the Huggingface [7] library, were fine-tuned. The models used were BERT-base-uncased [8], beto-emotion-analysis [9], BERT-base-spanish-cased [10], DeBERTa-base [11], RoBERTa-base [12] and XLNet-base-cased [13].

In order to prepare the data for analysis, several preprocessing methods were applied. This involved removing links, usernames, hashtags, and emojis from the text. By eliminating these elements, we aimed to focus solely on the textual content and minimize any potential noise or distractions.

Furthermore, to ensure optimal model training, a study of hyperparameters was conducted. This consisted in exploring different combinations of hyperparameters values to identify the most effective settings. To streamline this process, we leveraged the WandB tool, which provided valuable insights and facilitated efficient experimentation. Our objective was to optimize the performance and accuracy of our models by fine-tuning the hyperparameters, ultimately resulting in outcomes that are more robust and reliable.

After completing the training process and generating predictions for the individual messages, a user prediction methodology was employed for each task. In the case of Task 1a, once labels were predicted for each message of a user, a percentage-based approach was applied. The user's label will be determined by considering the majority label among the predictions generated for their individual messages. In other words, if a significant proportion of the user's messages are consistently predicted to belong to a particular class, that class will be assigned as the user's label. This approach ensures that the user's label reflects the prevailing classification consensus derived from their message predictions. For example, if a user has 20 messages and 14 of them were classified as 1, the user was automatically labelled as 1.

The user's label was derived by computing the average value of the predictions assigned to their individual messages. By aggregating the predicted values in this manner, we obtained a representative label that captures the overall tendency of the user's messages. This averaging process allowed us to capture the collective sentiment or characteristic exhibited by the user's messages and assign a label that reflects their overall classification.

Figure 1 shows a summary of the methodology used for Task 1.

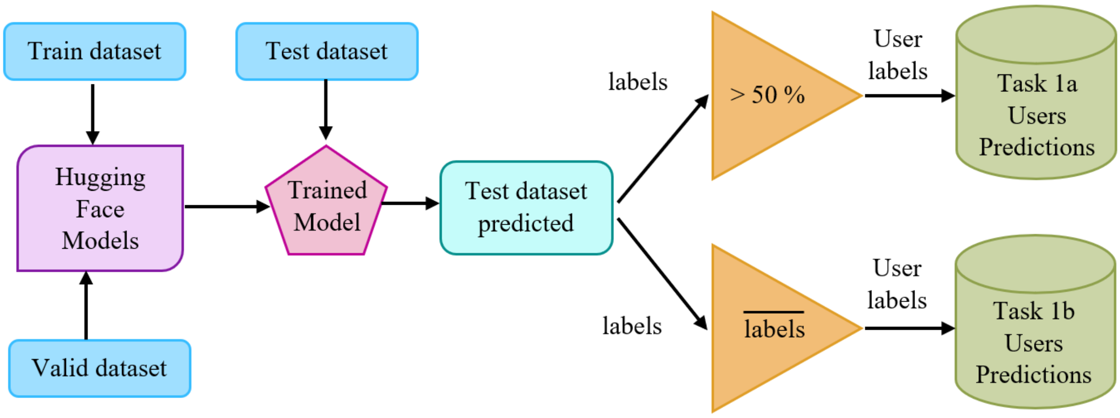

Figure 1: Methodology followed in Task 1

Table 7 shows the results achieved during the training phase for Task 1a using the models mentioned before.

Table 7
Experimental results Task 1a

	Messages Predictions		User Predictions	
Models	Accuracy	F1-score	Accuracy	F1-score
BERT-base-uncased	0.68	0.67	0.80	0.74
Beto-emotion-analysis	0.70	0.70	0.86	0.83
bert-base-spanish-wwm-cased	0.72	0.72	0.89	0.85
roberta-base-bne	0.72	0.70	0.86	0.80
xlnet-base-cased	0.67	0.66	0.80	0.72
deberta-v3-base	0.70	0.70	0.91	0.89

Once this first study was done, a new study of hyperparameters was carried out with the models that obtained the best results in the previous experiment. The hyperparameter search was done using the WandB tool with strategy *grid*. Table 8 shows the hyperparameters space used for this study. Each model was trained with these combinations of hyperparameters, and the best four combination values are shown in Table 9.

Table 8
Hyperparameters space

Hyperparameter	Values
Weight Decay	[0.1, 0.01]
Maximum Length	[128, 256]
Learning Rate	[5e-5]
Batch Size	[16,32]

Table 9
Best combination of models and hyperparameters

Models	Weight Decay	Max. Len.	Learning Rate	Batch Size
H1-roberta-base-bne	0.1	128	5e-5	32
H2-roberta-base-bne	0.01	128	5e-5	32
H3-deberta-v3-base	0.1	256	5e-5	32
H4- bert-base-spanish-wwm-cased	0.01	128	5e-5	32

Table 10 provides a comprehensive overview of the results obtained for Task 1a using the optimal hyperparameters. By carefully selecting and fine-tuning the hyperparameters, we aimed to maximize the performance and achieve the most accurate predictions for this specific task.

Table 10
Experimental results for Task 1a using the best hyperparameters

	Messages Predictions		User Predictions	
Models	Accuracy	F1-score	Accuracy	F1-score
H1-roberta-base-bne	**0.73**	**0.72**	**0.89**	**0.85**
H2-roberta-base-bne	**0.74**	**0.74**	**0.94**	**0.93**
H3-deberta-v3-base	0.69	0.68	0.89	0.85
H4- bert-base-spanish-wwm-cased	0.72	0.72	0.88	0.85

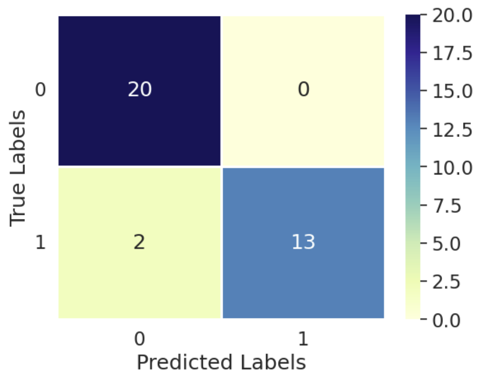

Figure 2: Confusion matrix of the best model *H2-roberta-base-bne* for Task 1a

In order to gain a deeper understanding of the failures, we conducted an analysis of the errors made by the models. Specifically, we delved into the results obtained from the best model. Figure 2 depicts the confusion matrix for *H2-roberta-base-bne* model when making user predictions. It is worth noting that the model exhibits superior performance in classifying users who do not have an eating disorder. A noteworthy observation is that 91% of users with an eating disorder were correctly classified, while users without an eating disorder were classified with 100% accuracy.

Upon examining the two users incorrectly labelled as 0 by the model, it becomes apparent that both individuals made comments pertaining to topics such as the “gym” and “food”, including references to “supplements”, “proteins”, “BCAAs”, and similar content. Such messages have the potential to confuse the model since not all individuals making such comments necessarily have an eating disorder. Consequently, correctly solving this task can prove challenging due to the nuanced nature of distinguishing between individuals who genuinely require assistance and those who do not solely based on such comments.

Taking into consideration the best model results in Task 1a experiments, the two models with highest results were trained to perform Task 1b. Results are shown in Table 11.

Table 11
Experimental results Task 1b

	Message Predictions		User Predictions	
Models	RMSE	Pearson	RMSE	Pearson
H1-roberta-base-bne	0.39	0.56	0.26	0.84
H2-roberta-base-bne	**0.39**	**0.58**	**0.25**	**0.86**

4.2. Task 2: Depression detection

To perform Task 2c, the model with best results in Task 1a was trained. The experiments done in this task are focused on *roberta-base-bne* model as it was the one that obtained best scores. The first phase of the study consisted of training the model mentioned before with the hyperparameter combination *H2*.

However, the results were not as expected, and another experiment was done. In addition to applying the above hyperparameters, a data augmentation method was applied to the train dataset to improve the results. Moreover, the training dataset has been restricted using only the three most representative classes: "control", "suffer+against" and "suffer+in favour". This technique is justified on the grounds that the "suffer+other" class could generate ambiguity during the model learning process, due to the similarity of the messages with the other two classes, "suffer+against" and "suffer+in favour".

The technique of back translation [14] was used with the minority classes, translating messages from Spanish to English, English to French and French to Spanish. Table 12 shows the distribution of the train dataset before and after the data augmentation method was applied.

Table 12
Distribution before and after applying data augmentation to the train dataset

	Control	Suffer+against	Suffer+in favour	Suffer+other
Before back translation	2275	1016	1062	97
After back translation	2275	**2124**	**2032**	**194**

Once the individual messages of each user were predicted, a method was employed to determine the class of each user. In this case, the classification of the user was determined by identifying the class that received the highest frequency of predictions. By selecting the class that was most commonly predicted across the user's messages, we aimed to assign a definitive label to each user that reflects the prevailing classification consensus. Figure 2 shows the methodology used in this task.

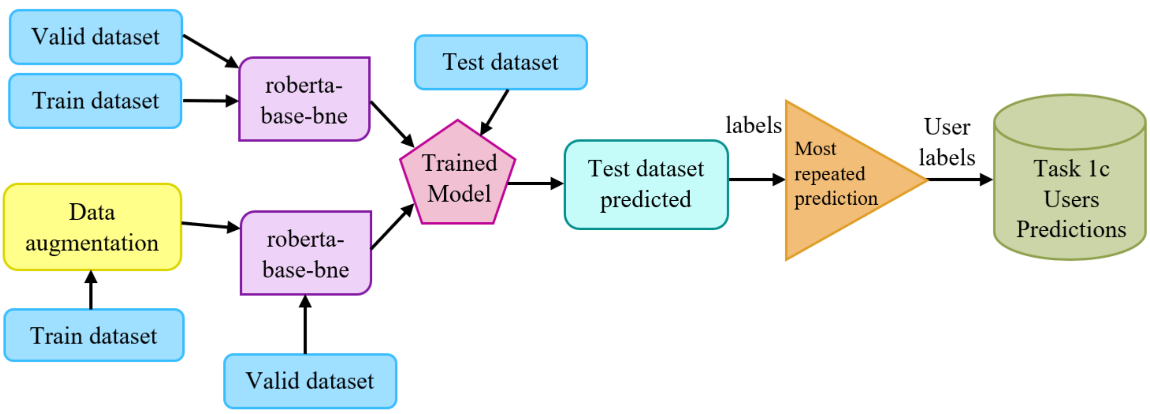

Figure 3: Methodology followed in Task 2

Results of both experiments are shown in Table 13, where Class 0 refers to "*Control*", class 1 "*Suffer+against*", class 2 "*Suffer+in favour*" and class 3 "Suffer+other". The scores did not meet our expectations and it could be improved in the future. The model *H2-roberta-base-bne* refers to the first experiment. The second technique used is represented by the model *H2-DA-roberta-base-bne*, where data augmentation was applied and the class "suffer+other" was ignored.

Table 13
Experimental results Task 2c

Models	F1 Class 0	F1 Class 1	F1 Class 2	F1 Class 3	Macro-F1
H2-roberta-base-bne	0.61	0.20	0.16	0.03	0.25
H2-DA-roberta-base-bne	0.67	0.17	0.25	0.00	0.25

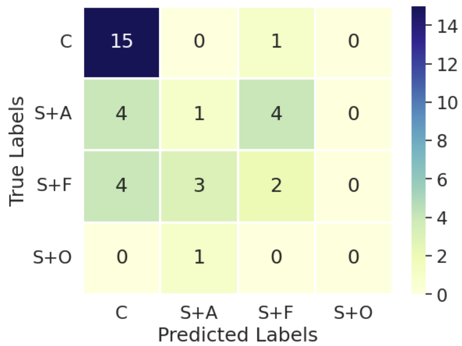

Figure 4 Confusion matrix of the best model *H2-DA-roberta-base-bne* for Task 2c

Figure 4 shows the confusion matrix for H2-DA-roberta-base-bne model when user predictions were made. Similar to Task 1a, the model demonstrates improved prediction accuracy for the majority class, which in this case is class 0 (Control), corresponding to label 0 in binary classification. Consequently, it can be inferred that our models exhibit stronger performance in predicting the negative class. Specifically, the "Control" class is correctly predicted at a rate of 94%, whereas less than 50% of the other classes are accurately predicted. As anticipated, class 3 (Suffer+other) is predicted incorrectly, primarily due to its absence during model training and the presence of only one user in the test dataset intended for model evaluation.

We have observed that the performance could depend on the number of instances available for each class. Therefore, it seems that increasing the number of instances for class 1 (suffer+against), class 2 (suffer+in favour) and class 3 (suffer+other) would improve its detection. In fact, data augmentation of these classes was carried out, but it seems that it has not been enough to ensure that these minority classes were correctly predicted.

5. Results

For the evaluation phase, model *H2-roberta-base-bne* was used for Task 1a and Task 1b and *H2-DA-roberta-base-bne* model was used to overcome Task 2c.

Our results [15] in the competition for Task 1a among the participants are shown in Table 14 and Table 15. The results were not as expected, and it will be improved in future works.

Table 14
Final results for Task 1a

Rank	Team	Run	Accuracy	Macro-P	Macro-R	Macro-F1
1	CIMAT-NLP-GTO	0	0.967	0.964	0.969	0.966
5	VICOM-nlp	2	0.880	0.878	0.885	0.879
19	**I2C-UHU**	**0**	**0.653**	**0.762**	**0.696**	**0.641**
22	UPM	2	0.453	0.719	0.523	0.349

Table 15
Final results for Task 1a (ERDE)

Rank	Team	Run	ERDE5	ERDE30	latencyTP	speed	Latency-weightedF1
1	CIMAT-NLP-GTO	0	0.334	0.018	6	0.898	0.863
5	Plncmm	0	0.498	0.074	10	0.817	0.679
16	**I2C-UHU**	**0**	**0.236**	**0.152**	**3**	**0.959**	**0.679**
22	Xabi IXA	2	0.325	0.237	3	0.959	0.589

Table 16 and Table 17 show the results of the competition of Task 1b among the participants. In this case, position 4 was achieved, so the strategies used to overcome Task 1a were more useful to carry out this part of the task.

Table 16
Final results for Task 1b

Rank	Team	Run	RMSE	Pearson_coefficient
	BaseLine-RoBERTa Base	2	0.178	0.906
4	**I2C-UHU**	**0**	**0.240**	**0.827**
10	UPM	1	0.324	0.586
17	Xabi IXA	2	0.503	0.352

Table 17
Final results for Task 1b

Rank	Team	Run	p@5	p@10	p@20	p@30
	BaseLine-RoBERTa Large	1	0.800	0.800	0.900	0.900
1	CIMAT-NLP	0	0.600	0.600	0.700	0.767
12	**I2C-UHU**	**0**	**1.000**	**0.700**	**0.750**	**0.700**
17	Xabi IXA	0	0.600	0.600	0.600	0.467

Our results in the competition of Task 2c can be seen in Table 18 and Table 19. Our best results were achieved in this task. A first and fourth position was achieved.

Table 18
Final results for Task 2c

Rank	Team	Run	Accuracy	Macro-P	Macro-R	Macro-F1
	BaseLine-RoBERTa Large	1	0.483	0.389	0.378	0.360
4	**I2C-UHU**	**0**	**0.315**	**0.307**	**0.253**	**0.232**
7	DepNLP UC3M GURUDASI	2	0.322	0.362	0.315	0.227
10	SPIN	2	0.248	0.434	0.292	0.161

Table 19
Final results for Task 2c (ERDE)

Rank	Team	Run	ERDE5	ERDE30	latencyTP	speed	Latency- weightedF1
	BaseLine-DeBERTa	0	0.330	0.190	2	0.984	0.695
1	**I2C-UHU**	**0**	**0.272**	**0.198**	**2**	**0.984**	**0.670**
7	plncmm	0	0.348	0.232	2	0.984	0.645
10	SPIN	2	0.431	0.245	3	0.967	0.609

The observed promising results can likely be attributed to the specific training technique employed. By focusing exclusively on the three majority classes ("control," "suffer+against," and "suffer+in favour") during the training phase, while bypassing the "suffer+other" class, it is plausible that the model achieved heightened effectiveness, leading to favorable outcomes. This selective approach to training allowed the model to concentrate its learning on distinguishing between the most prevalent classes, potentially contributing to the attainment of highly satisfactory results.

6. Conclusions

In this paper, we presented our proposal for MentalRiskES (Early detection of mental disorders risk in Spanish) and the results obtained in the shared Task for IberLEF 2023. Our approach consisted of fine-tuning transformer-based models using models from Hugging Face library. Moreover, a study of hyperparameters was carried out. Different approaches have been applied to each classifier to achieve the proper results. Our final model for Task 1a achieved a 0.641 Macro-F1 and got the nineteenth position in the ranking. For Task 1b, our model achieved a 0.240 RMSE and got the fourth position in the ranking. Finally, for Task 2c our model achieved a 0.232 Macro-F1 and got the fourth position in the ranking. It is worth mentioning the first place achieved in ERDE Multiclass Classification of Task 1c.

In future works we plan to explore other techniques of creating ensemble as well as do more exhaustive hyperparameters search for each classifier and apply this technique to make a real investigation of the early detection of mental disorder on Telegram.

7. References

[1] Guntuku, Sharath Chandra, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences* 18: 43-9.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30.

[3] Shorten, Connor, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data* 8: 1-34.

[4] Greco, Candida M., Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters* 167: 204-11.

[5] Wang, Shirley B. 2021. Machine learning to advance the prediction, prevention and treatment of eating disorders. *European Eating Disorders Review* 29 (5): 683-91.

[6] Benítez-Andrades, José Alberto, José Manuel Alija-Pérez, Isaías García-Rodríguez, Carmen Benavides, Héctor Alaiz-Moretón, Rafael Pastor Vargas, and María Teresa García-Ordás. 2021. BERT model-based approach for detecting categories of tweets in the field of eating disorders (ED). Paper presented at 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), .

[7] Face, Hugging. No title. in Retrieved from SBERT: https://huggingface. co/sentencetransformers [database online]. 2022Available from https://huggingface.co/.

[8] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.
[9] Pérez, Juan Manuel, Juan Carlos Giudici, and Franco Luque. 2021. Pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv Preprint arXiv:2106.09462*.
[10] Wu, Shijie, and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv Preprint arXiv:1904.09077*.
[11] He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv Preprint arXiv:2006.03654*.
[12] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv Preprint arXiv:1907.11692*.
[13] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems* 32.
[14] Sugiyama, Amane, and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. Paper presented at Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019).
[15] Mármol-Romero, Alba María and Moreno-Muñoz, Adrián and Plaza-del-Arco, Flor Miriam and Molina-González, María Dolores and Martín-Valdivia, María Teresa and Ureña-López, Luis Alfonso, and Montejo-Ráez Arturo. 2023. Overview of MentalRiskES at IberLEF 2023: Early detection of mental disorders risk in spanish. *Procesamiento Del Lenguaje Natural* 71.