SPIN at MentalRiskES 2023: Transformer-Based Model for Real-Life Depression Detection in Messaging Apps

Irune Zubiaga[1],*, Raquel Justo[1]

[1] *University of the Basque Country, Sarriena no number, 48940 Leioa, Spain*

Abstract

Depression is a prevalent and severe mental health condition that significantly impacts global population, causing personal suffering and reduced quality of life. Its symptoms are often visible on social media and digital platforms, making them valuable for detecting depression. This paper represents our submission for the MentalRiskEs task at IberLEF 2023. We present a novel hierarchical model for real-time chat applications, using natural language processing techniques to identify individuals at risk. Our approach combines similarity-based stance representation with a sentence-level transformer encoder block, reducing manual effort and time required for feature selection. Our focus includes binary classification of depressed and non-depressed users, as well as multi-class classification based on the user's coping mechanisms.

Keywords

Anchor sentences, Transformers, Depression Detection, Natural language

1. Introduction

Depression is a pervasive mental health disorder that affects millions of people worldwide, causing significant distress and impairing their daily functioning. It is characterized by persistent feelings of sadness, loss of interest or pleasure in activities, changes in appetite and sleep patterns, fatigue, difficulty concentrating, and thoughts of self-harm or suicide. This disorder does not only represent the leading cause of years lived with disability, but also contribute significantly to the global burden of suicide, which remains a major cause of death. Moreover, the economic impact of mental health conditions is substantial, surpassing the direct costs of care due to significant productivity losses [1]. Early detection and intervention are crucial to alleviate the suffering associated with depression and prevent its long-term consequences [2].

Traditionally, the diagnosis of depression has heavily relied on clinical interviews and self-report measures administered by mental health professionals. However, the rise of digital platforms, social media, and messaging applications has opened up new avenues for exploring innovative approaches to depression detection and monitoring.

In recent years, researchers have increasingly turned to natural language processing (NLP) and deep learning techniques to analyze textual data and detect mental health conditions, including depression, from user-generated content. These computational methods offer promising

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

irune.zubiaga@ehu.eus (I. Zubiaga); raquel.justo@ehu.eus (R. Justo)

0000-0002-3142-539X (R. Justo)

opportunities for automated and scalable approaches to identify individuals at risk or experiencing symptoms of depression [3]. NLP techniques such as topic modeling [4] and sentiment analysis [5] have been successfully used for depression detection in written text. Feature engineering, involving the manual crafting or extraction of relevant features, has been leveraged to detect depression, encompassing lexical, syntactic, and psychological markers [6]. Other commonly used approaches include the utilization of LIWC (Linguistic Inquiry and Word Count) features [7] and n-grams [8]. In the last few years transformer-based models have achieved state-of-the-art results in this task [9]. Ensemble learning methods have combined multiple models to enhance overall performance [10], while transfer learning techniques have utilized pre-trained models and fine-tuning to improve detection [11].

In this paper, we aim to contribute to the field of depression detection by presenting a hierarchical transformer based-model that, combined with similarity based features, identifies depression in text-based data. Specifically, this work focuses on the context of messaging applications.

The paper is organized as follows: Section 2 explains the task that was carried out and provides an insight of the corpus that was used to train the models. Section 3 outlines our methodology, explaining the data pre-processing and the used model architecture. Section 4 presents the experimental results and performance evaluation. Finally, Section 5 concludes the paper, summarizing the findings and discussing future directions for research.

2. Task and Corpus

This research focuses on early identification of depression in Spanish comments extracted from the messaging platform Telegram. Specifically, two tasks were carried out: Binary classification between depressed and non-depressed users and multi-class classification considering the user's way to cope with the affliction. For the multi-class classification task we will consider four classes: *control*, *suffer+in favor*, *suffer+against* and *suffer+other*. The *control* class will gather the users that show little to no signs of depression. The *suffer+against* class will gather individuals who are actively working towards overcoming depression while the *suffer+in favor* class will gather those who may feel overwhelmed by it and are not fighting against the illness. Lastly, the *suffer+other* class will gather users who openly discuss their depression without providing information about their efforts to combat it. In order to carry out this task, the MentalRiskES 2023 [12] corpus was provided. The corpus is divided into 3 subsets, each related to a different disorder. The target disorders encompass eating disorders, depression, and an undisclosed condition specifically included to evaluate the robustness of approaches for previously unknown disorders. The unknown disorder was anxiety. Table 1 shows example messages of users of each class.

Since our goal is to build a depression detection system, we were only provided with the MentalRiskES 2023 corpus regarding this affliction. This subset consists of messages collected from public Telegram groups that revolve around various topics directly related to this disorder and employ Spanish as the main language. It is noteworthy that a significant proportion of the messages within the corpus are written in Latin American Spanish, incorporating dialects from countries such as Argentina, Mexico, and others. These conversations involve hundreds of users that commonly employ informal language characterized by frequent typos, shortened

Table 1
Examples of user messages from the set regarding depression of the MentalRiskES 2023 corpus attributed to each of the classes. In the binary classification task all three "suffer" classes are clustered together under the positive class (the class that gathers depressed users).

Class	Original sentence	Translation
suffer + in favour	•Hola , estoy realmente mal , no se que hacer con mi vida •Porque las cosas no pueden acabar bien	•Hello, I'm really feeling down, I don't know what to do with my life •Because things can't end well
suffer + against	•Yo tengo depresion y soledad , pero salgo ahí fuera e intento apilar toda las particular de motivación que puede encontrar en el día para formar al menos una bola de tierra que pueda para seguir avanzando. •Hay que salir , hacer deporte , tirar las pastillas a la basura y producir las sustancias químicas que nuestro cerebro necesita.	•I have depression and loneliness, but I go out there and try to gather every bit of motivation I can find in the day to form at least a ball of soil that I can use to keep moving forward. •We need to go out, engage in sports, throw away the pills, and produce the chemical substances that our brain needs.
suffer + other	•yo estoy diagnosticado con depresion mayor •hay alguna juntada entre los argentinos del grupo?	•I have been diagnosed with major depression •is there any gathering among the Argentinians in the group?
control	•pareces más menor q yo •yo se bailar bachata mas menos	•you look more younger than me •i know how to dance bachata more or less

words, English expressions and the use of emoticons, reflecting typical practices observed in social media contexts. Some emoticons appeared as icons but most of them were in textual form: the description of the emoticons appeared instead of the icon (ex. cara sonriente/ smiley face). Some other emoticons were composed by ASCII symbols (ex. :)).

To ensure anonymity, the extracted messages underwent an anonymization process. Subsequently, a manual labeling process was conducted through the Prolific service [13], which facilitated the recruitment of annotators for online research. Each user's history was labeled by 10 annotators, and the probability of a disorder was determined based on the ratio of annotators who identified evidence of the targeted disorder to the total number of annotators (10). This measure enables regression analysis of the systems, allowing evaluation not solely based on a majority vote, but also considering how closely the prediction tool aligns with the confidence of a group of human judgments. The statistics of the corpus are shown in Table 2.

The challenge presents a unique online scenario, requiring the detection of potential risks within an ongoing stream of data. The evaluation process comprised multiple rounds of data

Table 2
Statistics of the depression related subset of the MentalRiskES corpus.

Partition	User number	Messages per User					
		Total messages	Mean	Max	Min	Std	Median
Train	175	6248	36	100	11	28	26
Validation	10	624	62	100	11	38	68
Test	149	5164	35	100	11	21	31
Total	334	12036					

testing, where each round featured a single message from every user. Initially, messages from all 149 users appeared in each round. However, in round 12, a gradual decline in user participation was observed, with some users ceasing to contribute. This pattern continued throughout the evaluation, resulting in a progressive reduction in the number of participating users. Ultimately, in the final round, only messages from 10 users were available for analysis. In each round, the messages from that specific round and all the previous rounds were utilized for prediction, ensuring that the model had access to a comprehensive history of user interactions. This approach allowed for a thorough assessment of the ability of the system to swiftly identify and address the issues at hand, even as user participation dwindled over time.

3. Methods and System Description

In order to tackle the task effectively, we devised two distinct systems. The first system utilized a transformer-based model for binary classification, enabling us to classify users as either depressed or non-depressed. Building upon this foundation, we further extended the model to address the subsequent challenge of determining the stance of users who exhibited positive indications of depression. By leveraging the inherited knowledge and representations from the binary classification model, our second system aimed to classify the user's stance based on their expressed sentiments and attitudes. In both cases the system incorporates a zero-shot approach that determines the membership of input sentences in each class by measuring their cosine similarity to a designated set of anchor sentences. These anchor sentences are chosen to represent each class. By calculating the cosine similarity, the system can assess the degree of resemblance between the input sentences and the anchor sentences, enabling classification even without prior training on specific examples. Before applying any model, the emoticons of the input that were not composed by ASCII symbols were converted to textual form.

3.1. Binary Classification Model

The binary classification model was trained by utilizing all the messages from each target user as inputs. The architecture, as depicted in Figure 1, begins by obtaining the message embeddings using a Sentence Transformer. Specifically, we used the pretrained model *paraphrase-spanish-distilroberta* [14]. Said model is designed as a sentence and short paragraph encoder, mapping input texts to a 768-dimensional dense vector space. It has been made by training a *bertin-*

Table 3
Example sentences of the corpus and their cosine similarities with the anchor sentence "Me siento deprimido" (I feel depressed).

Original sentence	Translation	Similarity
Ah bueno . La costumbre jaja	Oh well. The habit haha.	0.14
este grupo es como terapia ?	is this group like therapy?	0.21
últimamente he pensando en suicidarme	lately I have been thinking about killing myself	0.54

roberta-base-spanish model following a teacher-student transfer learning approach using parallel English-Spanish sentence pairs. The model captures semantic information and can be used for tasks such as information retrieval, clustering, and sentence similarity. The training process involved concatenating multiple datasets with sentence pairs in English and Spanish.

The embeddings are then passed through a Transformer encoder with two attention heads, allowing the model to automatically discern which inputs to focus on and which to disregard. Subsequently, the average of these embeddings is computed, resulting in a singular representation for each user. Simultaneously, the cosine similarity between the initial embeddings and the embedding of the anchor sentence "Me siento deprimido" (I feel depressed) is calculated, generating a total of n similarities (n being the number of user messages). The mean, standard deviation, and maximum of these similarity values are computed and concatenated with the user representation obtained from the transformer. This final representation, encompassing the user representation and the mean, standard deviation, and maximum of the similarities, is then fed into a multilayer perceptron (MLP) to produce a prediction for the user's class. The MLP consists of three layers: an input layer, a hidden layer with 60 neurons and an output layer. Batch normalization was applied to enhance training, and the ReLU activation function was used to capture nonlinear relationships effectively. The model was trained for 40 epochs with a learning rate of 2e-4 and a batch size of 16.

Additionally, in the test phase, we define a threshold for the cosine similarity in round one. If the calculated cosine similarity value exceeds 0.5, we classify the user as positive, indicating potential signs of depression. Conversely, if the cosine similarity value is below 0.5, we apply the model to get a prediction. This value does not normally exceed 0.5 unless it show clear symptoms of depression (see Table 3). This threshold-based classification allows for a initial determination of the user's status, enabling effective identification of individuals who may require further attention and support. This emphasis on prioritizing false positives over false negatives stems from the sensitive nature of depression, where proactive identification of potential cases takes precedence over waiting for definitive detection.

3.2. Multi-class Classification Model

The multi-class classification model builds upon the foundation of the binary classification model (see Section 3.1) to categorize users into four distinct classes: *control*, *suffer+against*, *suffer+in favour*, and *suffer+other*. The architecture of this model is depicted in Figure 1.

Table 4
Anchor sentences for the multi-class classification task.

Topic	Original sentence	Translation	Polarity
Overcome depression	Voy a superarlo	I will overcome this	Positive
	No creo poder superarlo	I don't think I'll be able to overcome it	Negative
Emotional state	Estoy contento	I am happy	Positive
	Estoy desanimado	I feel disheartened	Negative
Keep going	Saldré adelante	I will keep going	Positive
	Estoy hundido	I am feeling crushed	Negative

To accomplish this task, the binary classification system is initially employed to differentiate between depressed and non-depressed individuals. If a user is labeled as depressed, the second system is engaged; otherwise, the user is classified as control. In the second system, the initial embeddings and user representations obtained from the binary classification system are retrieved. Subsequently, the model computes the cosine similarity between the embeddings of the user messages and the set of anchor sentences outlined in Table 4. These sentences were thoughtfully selected to capture various aspects such as the user's perception of their ability to overcome depression, their emotional state, and their resilience or feelings of defeat.

After obtaining the sets of similarity values for each sentence, which resulted in 6 sets of n similarities, an equation is applied for each topic, as illustrated in Figure 1. The equation $s_{neutral} = 1 - \frac{(s_{positive} + s_{negative})}{2}$ is used to derive a value that lies in the middle range, enhancing the detection of the *suffer + other* class. As a result, 9 sets of n similarities are generated. Subsequently, the mean, maximum, standard deviation and median of the similarities are calculated resulting on 4 vectors of dimension 1x9. These vectors are then concatenated with the user representation and fed into a MLP that predicts the user's stance. As in the binary classification model, the MLP used for this task consists of an input layer, a hidden layer with 60 neurons and an output layer. Batch normalization was applied and ReLU was used as the activation function.

The model was once again trained for 40 epochs with a learning rate of 2e-4 and a batch size of 16.

4. Results

The model was evaluated on the test set of the corpus using the evaluation functions provided in the MentalRiskEs GitHub repository [15]. As described in Section 2, the test set was divided into 100 rounds, with a maximum of one message per user in each. The evaluation process followed two criteria. Firstly, the overall performance of the system was assessed by iterating through the rounds and recording the obtained metrics at each one. With this evaluation approach we evaluate the user's state in each round, allowing for the prediction to change from one round to another. This way it becomes possible to analyze potential signs of depression that may emerge during conversations and are of temporary nature, such as those associated with the

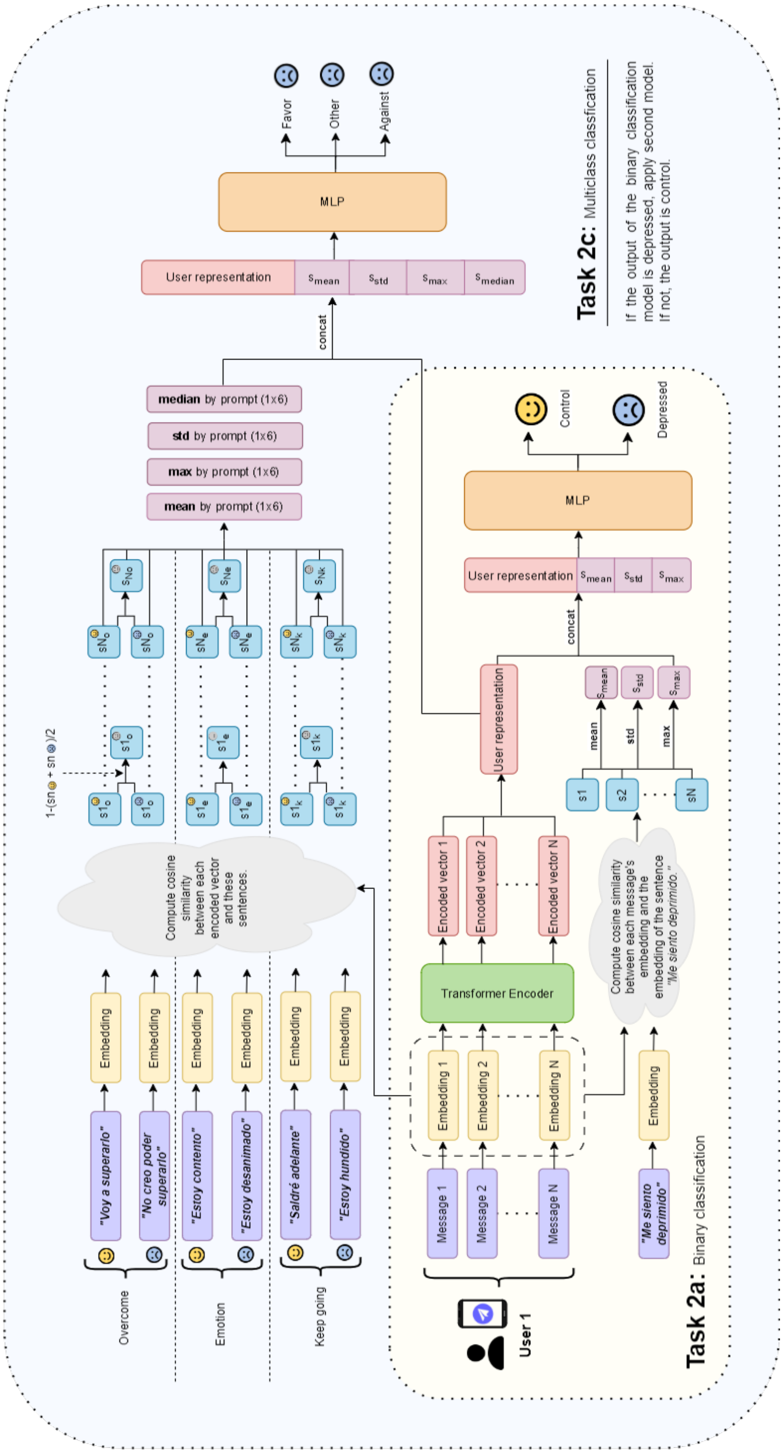
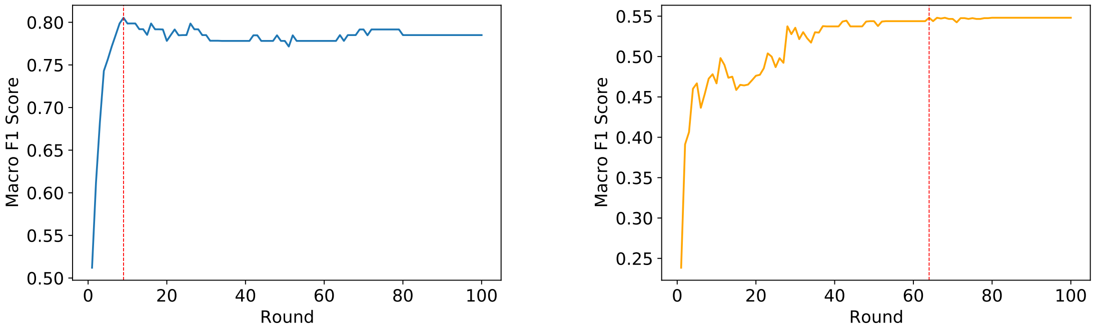


Figure 1: Architecture of the models that were used to carry out the binary classification and multi-class classification tasks.

loss of a loved one, significant life setbacks, or challenging life circumstances. Secondly, an approach focused on early detection was employed: once a user was evaluated as positive, their state predictions were no longer updated. For instance, if a user was labeled as positive in round 12, they would not be reevaluated in subsequent rounds, and their label would remain consistent. This ensured that the diagnosis became terminal, regardless of whether the positive identification was attributed to a transient feeling of sadness. This approach was consistent with the evaluators' strategy for the task.

Considering the first evaluation approach, this is, by taking into account the results of each round without imposing significant limitations, we get the results gathered in Table 5. Additionally, a plot that shows the evolution of the F1 Macro score is provided (see Figure). In Table 5 the results of the most relevant rounds for each task were selected. The selection was carried out considering the evolution of the system's performance shown in Figure 2. This provides a comprehensive overview of the behaviour of the system.



(a) Binary Macro F1-score

(b) Multiclass Macro F1-score

Figure 2: Evolution of the F1 Macro score through the rounds for each task.

As observed, the F1 score exhibits a gradual increase in value with each round, starting from a relatively low point. This upward trend continues until it reaches a stability point where further data acquisition ceases to significantly impact the outcome. For the binary classification task, this stability point is observed around round 9, while for the multi-class classification task, it occurs around round 40. It is noteworthy that the binary classification task achieved a substantial level of precision, despite utilizing a relatively small number of messages (specifically, 9 messages per user). This highlights the model's capacity to leverage limited data and achieve commendable results.

Finally, when accounting for the exclusion of evaluations that initially classified a subject as positive but were later classified as negative in subsequent rounds, we obtained the results in Table 6. Table 7 displays the results that were actually sent to the evaluation server. Although the exact cause of the discrepancy between the two sets of results remains unclear, we believe it may be attributed to an error we made while sending the predictions to the evaluation server. Both the binary evaluation metrics and the latency based metrics are provided. The latency-based metrics remain consistent for both tasks. This is due to the multi-class classification model inheriting from the binary classification model, as these metrics focus on the detection

Table 5
Classification results in each round without imposing significant limitations.

(a) Results of the Binary Classification Task.

Round	Accuracy	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
1	0.61	0.72	0.58	0.51	0.61	0.61	0.61
2	0.62	0.61	0.61	0.61	0.62	0.62	0.62
3	0.68	0.68	0.68	0.68	0.68	0.68	0.68
4	0.74	0.74	0.74	0.74	0.74	0.74	0.74
5	0.76	0.76	0.76	0.76	0.76	0.76	0.76
6	0.77	0.77	0.77	0.77	0.77	0.77	0.77
7	0.79	0.79	0.79	0.79	0.79	0.79	0.79
8	0.80	0.81	0.81	0.80	0.80	0.80	0.80
9	0.81	0.81	0.81	0.81	0.81	0.81	0.81
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
100	0.79	0.80	0.79	0.78	0.79	0.79	0.79

(b) Results of the Multi-class Classification Task.

Round	Accuracy	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
1	0.56	0.24	0.28	0.24	0.56	0.56	0.56
2	0.52	0.41	0.39	0.39	0.52	0.52	0.52
3	0.57	0.41	0.40	0.41	0.57	0.57	0.57
4	0.64	0.47	0.45	0.46	0.64	0.64	0.64
5	0.65	0.47	0.46	0.47	0.65	0.65	0.65
6	0.62	0.44	0.44	0.44	0.62	0.62	0.62
7	0.64	0.45	0.46	0.45	0.64	0.64	0.64
8	0.65	0.47	0.48	0.47	0.65	0.65	0.65
9	0.66	0.47	0.49	0.48	0.66	0.66	0.66
10	0.64	0.47	0.47	0.47	0.64	0.64	0.64
11	0.64	0.49	0.51	0.50	0.64	0.64	0.64
12	0.66	0.48	0.49	0.49	0.66	0.66	0.66
13	0.64	0.47	0.48	0.47	0.64	0.64	0.64
14	0.64	0.47	0.48	0.47	0.64	0.64	0.64
15	0.63	0.46	0.46	0.46	0.63	0.63	0.63
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
39	0.66	0.54	0.56	0.54	0.66	0.66	0.66
40	0.66	0.54	0.56	0.54	0.66	0.66	0.66
41	0.66	0.54	0.56	0.54	0.66	0.66	0.66
42	0.67	0.55	0.57	0.54	0.67	0.67	0.67
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
100	0.68	0.55	0.57	0.55	0.68	0.68	0.68

of positives without differentiating between positive classes. As may be seen, the F1 score is considerably lower when following this evaluation approach, which can be attributed to the limitations we impose on the model: we do not allow it to adapt its answers with new information once a positive prediction is made.

Table 6
Early detection results of the Classification Task.

(a) Binary Evaluation Metrics.

Classes	Accuracy	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
2	0.685	0.771	0.708	0.671	0.685	0.685	0.685
4	0.483	0.448	0.440	0.393	0.483	0.483	0.483

(b) Latency Based Metrics.

Classes	ERDE5	ERDE30	Median Latency	Speed	Latency-weighted F1
2 & 4	0.298	0.151	3.0	0.982	0.724

Table 7
Early detection results of the Classification Task presented in the MentalRiskEs challenge.

(a) Binary Evaluation Metrics.

Classes	Accuracy	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
2	0.470	0.731	0.512	0.340	0.470	0.470	0.470
4	0.262	0.412	0.343	0.219	0.262	0.262	0.262

(b) Latency Based Metrics.

Classes	ERDE5	ERDE30	Median Latency	Speed	Latency-weighted F1
2 & 4	0.402	0.242	3.0	0.967	0.612

The CodeCarbon [16] library was used to track the emissions and energy consumption of the models. Each model used around 5e-5 KwH and emitted around 8e-6 g of CO2 per round.

5. Error analysis

In the multi-class classification task, one notable observation was that the *suffer + other* class was frequently not recognized with the same level of accuracy as the *suffer + against* and *suffer + in favor* classes. This discrepancy can be attributed to the relatively limited number of samples available for the *suffer + other* class compared to the other classes. With a smaller sample size, the classifier had difficulty accurately identifying instances belonging to the *suffer + other* class. In future studies, it would be valuable to explore oversampling methods to address the class imbalance issue and potentially improve the recognition of the *suffer + other* class. Additionally, expanding the pool of reference sentences specific to the *suffer + other* class could enhance the classifier's ability to discriminate and correctly classify instances within this category. Such advancements would contribute to a more comprehensive understanding of the nuances in multi-class classification for this particular task.

6. Conclusions and Future Work

In this paper, we proposed and analyzed a hierarchical transformer-based model for depression detection in real-life chat applications that leverages anchor sentences to enhance its predictions. Our model achieved a Macro F1 score of **0.78** for the binary classification task and **0.55** for the multi-class classification task and a **0.67** for the binary classification task and **0.39** for the multi-class task in early detection. These results highlight the effectiveness of our approach in capturing and utilizing contextual information for accurate classification.

Moving forward, we plan to further explore and optimize the model by expanding the evaluation to a larger test set and conducting in-depth analysis of various factors such as the effects of anchor sentences and hyperparameters like the number of attention heads. Additionally, we aim to investigate zero-shot approaches based on similarities. Energy usage and CO2 emission of the models is also to be further analyzed, considering the importance of this aspect.

References

[1] W. H. Organization, et al., World mental health report: transforming mental health for all (2022).

[2] F. Cacheda, D. Fernandez, F. J. Novoa, V. Carneiro, Early detection of depression: Social network analysis and random forest techniques, Journal of Medical Internet Research 21 (2019) e12554. doi:10.2196/12554.

[3] M. L. Joshi, N. Kanoongo, Depression detection using emotional artificial intelligence and machine learning: A closer review, Materials Today: Proceedings 58 (2022) 217–226. doi:https://doi.org/10.1016/j.matpr.2022.01.467, international Conference on Artificial Intelligence Energy Systems.

[4] P. J. Franz, E. C. Nook, P. Mair, M. K. Nock, Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform, Suicide and Life-Threatening Behavior 50 (2020) 5–18.

[5] N. V. Babu, E. G. M. Kanaga, Sentiment analysis in social media data for depression detection using artificial intelligence: a review, SN Computer Science 3 (2022) 1–20.

[6] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, G. Parker, Interpretation of depression detection models via feature selection methods, IEEE transactions on affective computing (2020).

[7] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of language and social psychology 29 (2010) 24–54.

[8] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, τ-ss3: a text classifier with dynamic

n-grams for early risk detection over text streams, Pattern Recognition Letters 138 (2020) 130–137.

[9] A.-M. Bucur, A. Cosma, P. Rosso, L. P. Dinu, It's just a matter of time: Detecting depression with time-enriched multimodal transformers, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 200–215.

[10] W.-Y. Wang, Y.-C. Tang, W.-W. Du, W.-C. Peng, NYCU_TWD@LT-EDI-ACL2022: Ensemble models with VADER and contrastive learning for detecting signs of depression from social media, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 136–139. doi:10.18653/v1/2022.ltedi-1.15.

[11] R. Poświata, M. Perełkiewicz, OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 276–282. doi:10.18653/v1/2022.ltedi-1.40.

[12] 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 2023: Early Detection of Mental Disorders Risk in Spanish, Procesamiento del Lenguaje Natural 71 (2023).

[13] Prolific, 2023. URL: https://www.prolific.co/, accessed: 6 12, 2023.

[14] A. Pérez, E. T. Ariza, L. G. Pinto, M. Mazuecos, hackathon-pln-es/paraphrase-spanish-distilroberta, Hugging Face Model Hub, 2021. URL: https://huggingface.co/hackathon-pln-es/paraphrase-spanish-distilroberta.

[15] SINAI-UJA, MentalRiskEs, GitHub repository, 2023. URL: https://github.com/sinai-uja/MentalRiskEs.

[16] Codecarbon, 2021. URL: https://codecarbon.io/.