NLP-UNED at MentalRiskES 2023: Approximate Nearest Neighbors for Identifying Health Disorders

Hermenegildo Fabregat[1,3], Andres Duque[1,2,*], Lourdes Araujo[1,2] and Juan Martinez-Romo[1,2]

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain
[2]IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain
[3]Avature Machine Learning, Spain

Abstract

This paper describes our participation in the first edition of the MentalRiskES Workshop (Early detection of mental disorders risk in Spanish) of IberLEF 2023 shared evaluation campaign, devoted to the early detection of different health disorders in Spanish comments from telegram users. As in other similar tasks, the original dataset is annotated at user level, that is why the proposed approach makes use of an Approximate Nearest Neighbors (ANN) based relabeling process in order to produce a message-level annotated dataset. This process has been validated in other contexts by analyzing English messages and in this case we try to cover the context described by this new Workshop by studying Spanish messages. To address the classification of new messages, the proposed approach analyzes their similarity to the relabeled dataset. Analyzing the results obtained in all the tasks from this workshop, our system obtains an interesting average result, highlighting the results obtained in Task 2.C where our system is placed among the top three participants.

Keywords

Mental Risk Detection, Approximate Nearest Neighbors, Social Media

1. Introduction

For individuals and society, mental disorders remain a major challenge. Early identification and intervention are critical to the effective management of these disorders and the mitigation of their potential impact on the well-being of individuals. Recent advances in technology and the increasing popularity of social media platforms have provided new opportunities to explore innovative approaches to detecting early signs of mental health disorders [1]. One of such platforms, Telegram, a widely used instant messaging application, offers a unique opportunity to study users' messages and potentially obtain valuable insights into their mental well-being.

A promising avenue for researchers to access large and diverse populations is the amount of data generated by social media users. Telegram, with its large user base and rich messaging fea-

IberLEF 2023, September 2023, Jaén, Spain
*Corresponding author.
gildo.fabregat@gmail.com (H. Fabregat); aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)
0000-0001-9820-2150 (H. Fabregat); 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

tures, allows individuals to openly express what they think, feel and experience. Consequently, mining and analysing Telegram messages may provide valuable indicators to help identify early risk factors for mental disorders.

In this paper we present our participation in the MentalRiskES task of the IberLEF 2023 shared evaluation campaign, which will be held in the context of the XXXIX Congress of the Spanish Society for Natural Language Processing (SEPLN 2023). Through the analysis of messages written in Spanish, the systems that take part in the task must be able to detect, as early as possible, the possible risk that certain users of Telegram may suffer from some kind of mental disorder. The task organisers provide an evaluation framework that measures, besides the systems' accuracy, the amount of information required to generate the response. The proposed system is based on the use of approximate nearest neighbor (ANN) techniques. Such approaches are well suited to potentially large data environments as they are designed to provide sub-optimal decision spaces in a reasonable amount of time. Furthermore, the proposed system uses an evidence-based approach, where for each decision made, the set of textual evidence that led to that decision can be provided. Finally, a corpus re-labeling method is proposed that minimises the possible noise derived from coarse-grained annotations. Having previously analysed the value of this approach by studying tasks focused on social networks such as Reddit [1, 2], this time the objective is to study the performance of this proposal for languages other than English and for new types of mental disorder.

The rest of the paper is structured as follows: Section 2 is devoted to explore existing systems performing early detection of different kind of disorders through textual data analysis. The main details of the task addressed in this paper are described in the Section 3. The developed system is described in Section 4 and the obtained results in Section 5. Finally, Section 6 offers some conclusions regarding this research and possible lines of work to be followed in the future.

2. Related Work

As a result of their plural and widespread use, social networks are increasingly becoming a field of study for the identification of users at risk of suffering from any kind of medical disorder. While works such as [3] and [1, 4, 2] are focused on the study of platforms such as Twitter or Reddit, other platforms such as Telegram, which could be considered as an instant messaging tool, have not been explored extensively to date. Although aspects of Telegram groups may be similar to the way Reddit works, Telegram groups enhance the feeling of dynamism as well as the sense of belonging to a collective. In order to study different types of disorders, the task organizers extracted messages written in Spanish from different groups related to the following disorders: anorexia, depression and anxiety.

Screening for different disorders in social network users has been approached using a large number of techniques, but Transformer-based approaches, such as [5] and [6], currently are the leaders in performance. To address the identification of at-risk users, both works propose Roberta-based classifiers. The performance of this type of classifiers in tasks as different as the identification of self-harm disorders and pathological gambling disorders is outstanding. [5] propose the use of classifiers based on XLM-RoBERTa-base [7] and for the sequential processing of a message history, they propose a decision function where they consider that in order to make

a decision, a window of N messages must be considered (retention). Then, after processing more than N messages, the window must be moved (forgetting), so they eliminate the old messages and include the most recent ones. Using a similar message window concept, [6] propose the use of a Roberta-large [8] based classifier exploring the feature space generated by the concatenation of text embeddings and a lexical metric feature space, covering aspects such as volumetry, lexical diversity, complexity and emotions. Unlike them, we propose a simple approach based on nearest-neighbor retrieval techniques. Given the sheer volume of information generated by social media platforms, we use ANN techniques. They provide good performance in processing large volumes of data, allowing scalable and efficient solutions.

On the other hand, for identifying a user at potential risk of suffering from some kind of disorder, approaches based on deep learning have always suffered from a great lack of explainability [9, 10]. However, there are opposing positions that offer positive and interesting results by analyzing the activation of different types of attention layers while processing certain message history [11]. In contrast to these systems, our approach allows for the direct extraction of textual evidences for the generated scores. These evidences can be useful for further manual analysis.

The proposed model was previously considered for the detection of users at risk of suffering pathological gambling disorders [12] and for the detection of users at risk of suicide [13]. The results obtained motivate us to continue the study of this approach, being this workshop an opportunity to address the analysis of different disorders and languages other than English.

3. MentalRiskES Task

MentalRiskES is a novel task on early risk identification of health disorders in Spanish comments from Telegram users. In this task, participating systems are asked to determine an individual's potential risk of suffering from certain health disorders. Performance depends not only on the accuracy of the systems, but also on how quickly the problem is detected. In this edition, the disorders considered by the organizers were eating disorders, depression and anxiety.

Task1 Detection of eating disorders (anorexia or bulimia):

Task1.a Binary classification: Determine if the user is affected or in control.

Task1.b Simple regression: Assess the probability of the user having anorexia or bulimia.

Task2 Detection of depression:

Task2.a Binary classification: Determine if the user is affected or in control.

Task2.b Simple regression: Assess the probability of the user experiencing depression.

Task2.c Multi-class classification: Categorize the user into one of four labels: "affected+against", "affected+in favor", "affected+other", or "control".

Task2.d Multi-output regression: Provide a probability for each of the above classes, indicating the likelihood of the user belonging to that class.

Task3 Detection of an unknown (anxiety) disorder:

Task3.a Binary classification: Determine if the user is affected or in control.

Task3.b Simple regression: Assess the probability of the user experiencing anxiety.

Additional information about the task can be found in [14].

3.1. Evaluation Metrics

The evaluation framework proposed by the task organisers covers the following aspects:

Performance: For both classification and regression tasks, a typical evaluation framework is proposed where metrics such as Precision, Recall, F1 and RMSE are reported.

Time: For all tasks, the proposed framework includes the wall-time required to process the entire dataset. Also, for the classification tasks, some metrics focused on the analysis of the number of messages needed to send an alert in case of detection of a positive user are proposed (Latency True Positive, Speed and F1-weighted).

Energy efficiency: The organisers also proposed the study of a set of metrics to analyse the carbon footprint generated by each participating system. Participants were required to use the Python library CodeCarbon [15] to obtain these metrics.

4. Proposed System

The model proposed by [12], which addresses the task of classifying messages in social networks through the use of Approximate Nearest Neighbors (ANN) techniques, has been employed. This type of approaches has proven to be very useful for processing large data collections, being this aspect critical in contexts related to social media data. At the same time, this kind of evidence-based approach offers the possibility of future manual analysis by providing potential evidence on the decisions predicted. The following sections describe the main components of the proposed model.

4.1. Data representation

We use an encoder from the Universal Sentence Encoder family [16] for encoding the text of each message in an N-dimensional vector space. Such models are trained and optimized for encoding texts longer than words e.g., sentences, phrases or short paragraphs. In short, for each message encoded by this model, a 512-dimensional vector is generated. In this work, we use the multilingual model based on Transformer [17], which supports the Spanish language. We do not perform any kind of fine tuning and we use a low-profile GPU (NVIDIA GeForce RTX 2060 Max-Q) in order to reduce execution time without increasing too much the energy consuming footprint.

4.2. Approximate Nearest Neighbors

The large amount of data generated daily in social networks makes the use of exact approaches for nearest neighbor estimation practically intractable. Unlike traditional nearest neighbor approaches, approximate techniques seek to deal with limitations derived from the amount of data to be explored by ensuring sub-optimal decision spaces in a reasonable amount of time. Currently there are different tools and approaches that have proven to be very successful when analysing recall results and queries per second [18]. These techniques collect different types of approaches, but most of them seek to make subdivisions in the representation space in order to generate a navigable space e.g., trees, being this the case of the Annoy library [19]. This library uses tree-like structures for the representation of nodes and random projections for the division of the subspace between adjacent nodes. We use this tool for retrieving nearest neighbours by exploring the cosine distance between new messages and the elements that are part of the Annoy space.

4.3. Relabeling process

In the same way as in tasks such as [1, 20], the task organizers only provided labels for the users, i.e. all messages from a user have the same label in relation to his possible health condition. In order to address this task with approaches that make use of knowledge extracted during the training, we designed a data relabeling approach. This approach attempts to reduce the noise inherent in considering the whole message history of a particular user under the same label. After generating the navigable index with the training data, we use an iterative approach to re-label the training data. Initially, we consider the original tags of each user to label their messages. At each iteration of the process, we check the label associated with a positive message by exploring the K nearest neighbors. To consider a message as positive, at least J positive messages must be retrieved, where $J <= K$. This approach is repeated until the training dataset converges to a set of messages and labels, where the relationship between messages that have the same label is maximized. Since the content of the messages constituting the navigable index is not changed, both the index and the final set of relabeled data are considered as the evidence pool to be used during the inference/classification of new messages during the testing phase. During relabeling, the following list of parameters was used depending on the task addressed: **Task1:** $J = 6$ and $K = 10$; **Task2:** $J = 4$ and $K = 10$; **Task3:** $J = 4$ and $K = 10$. We explored a subset from the training set to explore all the parameters required by the model. Section 5 includes all details about data partitions explored.

4.4. Tag and scoring function

After generating the navigable index, we modify the original tagging and scoring functions to adapt them to the different tasks considered. For processing each message M from a user X we use the following functions:

Binary classification We extract the K nearest neighbors and their associated labels. We determine that a user is positive if at least J retrieved messages are positive and the average distance is lower than 1.

Table 1
Results obtained by participating systems for Task 1.A (Selection and ranking criterion used: macro-f1). The best results are in bold.

Rank	Team	Run	Accuracy	Macro_F1	ERDE5	ERDE30	latencyTP	speed	latency-weightedF1
1	CIMAT-NLP-GTO	0	**0.967**	**0.966**	0.334	**0.018**	6	0.898	**0.863**
2	VICOM-nlp	2	0.880	0.879	0.169	0.070	3	0.959	0.832
3	VICOM-nlp	1	0.860	0.859	0.223	0.085	3	0.959	0.814
4	VICOM-nlp	0	0.853	0.850	0.226	0.111	3	0.959	0.794
5	BaseLine - Roberta Large	1	0.813	0.813	**0.163**	0.099	**2**	**0.979**	0.792
6	UNSL	1	0.913	0.913	0.433	0.045	8	0.857	0.776
7	CIMAT-NLP-GTO	1	0.847	0.847	0.379	0.065	6	0.898	0.761
8	CIMAT-NLP	1	0.820	0.820	0.370	0.088	5	0.918	0.752
9	BaseLine - Deberta	0	0.813	0.813	0.310	0.083	5	0.918	0.751
12	BaseLine - Roberta Base	2	0.700	0.694	0.186	0.132	**2**	**0.979**	0.722
11	NLP-UNED (Original)	0	0.760	0.760	0.268	0.118	3	0.959	0.738
18	NLP-UNED (Relabel)	1	0.760	0.749	0.303	0.196	3	0.959	0.666

Multiclass classification Since the model was originally intended for binary classification between at-risk and non-at-risk user, to process the multi-class scenario we first predict whether a user is positive or negative and then, for those positive cases we assign the retrieved majority class.

Regression If a new message is considered as positive in any of the previous scenarios, we calculate the score as follows: $scoring = 0.5 * (2 - D)$, where D is the average distance of the recovered neighbors.

During inference, the following list of parameters was used depending on the task addressed: **Task1:** $J = 8$ and $K = 10$; **Task2:** $J = 5$ and $K = 10$; **Task3:** $J = 8$ and $K = 10$. All these parameters were explored during the validation stage.

5. Results

In this section the main results obtained by our system during the test phase are presented. Also, in order to contextualise these results, we also include the results obtained by a selection of the best participating systems.

The organisers allowed the submission of up to three different runs for each subtask. We only sent two runs alternating the use of the relabelling method (i.e., Run0 does not make use of the relabelling process and Run1 implies relabeling the original message). The rest of the parameters were tuned testing the performance of the model during training. We divide the training set into training (70%) and validation (30%).

5.1. Binary classification

For Task 1.A, Table 1 shows the results obtained by the participating systems. Although the proposed system obtains good results in terms of latency, the accuracy results are considerably lower than those obtained by other participating systems. Unlike in the rest of the classification tasks, the deterioration produced after the application of the relabelling method stands out.

Table 2
Results obtained by participating systems for Task 2.A (Selection and ranking criterion used: macro-f1). The best results are in bold.

Rank	Team	Run	Accuracy	Macro_F1	ERDE5	ERDE30	latencyTP	speed	latency-weightedF1
1	UMUTeam	0	**0.738**	**0.737**	0.548	0.358	30.000	0.560	0.421
2	UNSL	1	**0.738**	0.733	0.567	0.148	14.000	0.791	0.609
3	UNSL	0	0.732	0.731	0.551	0.188	14.000	0.791	0.591
4	TextualTherapists	1	0.732	0.729	0.421	0.161	7.000	0.903	0.682
5	SINAI-SELA	0	0.725	0.720	0.395	**0.140**	4.000	0.951	**0.720**
6	UMUTeam	1	0.705	0.705	0.548	0.371	30.000	0.560	0.398
7	BaseLine - Roberta Large	1	0.698	0.690	0.290	0.159	4.000	0.951	0.704
8	SINAI-SELA	1	0.685	0.675	0.389	0.159	4.000	0.951	0.696
9	TextualTherapists	0	0.664	0.651	0.342	0.168	3.000	0.967	0.696
12	BaseLine - Deberta	0	0.664	0.642	0.303	0.153	**2.000**	**0.984**	0.719
10	NLP-UNED (Relabel)	1	0.651	0.648	0.411	0.207	6.000	**0.919**	0.624
16	NLP-UNED (Original)	0	0.624	0.617	0.404	0.212	5.000	**0.935**	0.627

On the other hand, the system seems to obtain a better performance-latency ratio in the tasks 2.A and 3.A, as shown in Tables 2 and 3. The performance improvements obtained by the relabelling method seem to have almost no negative impact on the latency of the system. Moreover, in overall terms, the results obtained show that our system obtains quite competitive results in comparison with other more complex proposals, as is the case of the baselines proposed by the organisers.

Table 3
Results obtained by participating systems for Task 3.A (Selection and ranking criterion used: macro-f1). The best results are in bold.

Rank	Team	Run	Accuracy	Macro_F1	ERDE5	ERDE30	latencyTP	speed	latency-weightedF1
1	CIMAT-NLP-GTO	2	**0.773**	**0.740**	0.691	0.188	7.000	0.908	0.757
2	BaseLine - Deberta	0	0.760	0.693	0.347	**0.165**	4.000	0.954	0.798
4	BaseLine - Roberta Large	1	0.720	0.630	0.324	0.179	**2.000**	**0.985**	**0.800**
5	CIMAT-NLP	0	0.673	0.614	0.769	0.250	14.000	0.802	**0.614**
7	CIMAT-NLP-GTO	0	0.633	0.593	0.710	0.283	7.000	0.908	0.654
8	BaseLine - Roberta Base	2	0.680	0.553	**0.309**	0.210	**2.000**	**0.985**	0.779
3	NLP-UNED (Relabel)	1	0.680	0.650	0.632	0.285	8.000	0.893	0.672
6	NLP-UNED (Original)	0	0.640	0.595	0.652	0.310	8.000	0.893	0.652

5.2. Multi-class classification

We tackle multi-class classification following the approach we described in Section 4.4. Analysing the results shown in Table 4, and taking into account the *latency-weightedF1* results to analyse the performance of the first phase, the results obtained by our system seem to be quite consistent with those obtained in Task 2.A, in contrast to the results obtained by the proposed baseline based on Roberta Large [8]. Finally, in terms of the selection of the final label, our system is one of the best proposals, i.e. based on the ranking offered by the organisers (macro-F1).

Table 4
Results obtained by participating systems for Task 2.C (Selection and ranking criterion used: macro-f1). The best results are in bold.

Rank	Team	Run	Accuracy	Macro_F1	ERDE5	ERDE30	latencyTP	speed	latency-weightedF1
1	BaseLine - Roberta Large	1	0.483	**0.360**	0.283	0.232	**2.000**	**0.984**	0.652
4	BaseLine - Deberta	0	0.456	0.293	0.330	**0.190**	**2.000**	**0.984**	**0.695**
5	plncmm	0	0.383	0.288	0.348	0.232	**2.000**	**0.984**	0.645
6	BaseLine - Roberta Base	2	0.356	0.274	0.307	**0.206**	**2.000**	**0.984**	0.659
7	I2C-UHU	0	0.315	0.232	**0.272**	0.198	**2.000**	**0.984**	0.670
2	NLP-UNED (Relabel)	1	**0.490**	0.358	**0.412**	0.203	5.000	**0.935**	0.638
3	NLP-UNED (Original)	0	0.450	0.339	0.408	0.211	5.000	0.935	0.627

5.3. Regression tasks

On the subtasks related to regression-based evaluations, the approach has not obtained the expected results, being in different occasions among the last positions of the ranking (criterion: RMSE). This may be due to the conditional generation of the score, where non-zero scores are only generated for those messages detected as positive.

5.4. Other evaluations

Since its conception, the model has been designed taking into account the processing time required per query. In contrast to other works [13], we have used an encoder based on Transformer. We tried to reduce any additional delay by using a GPU enabled computer. For energy efficiency reasons, we decided to use a low-profile GPU. As shown in Table 5, these decisions did not lead to a major deterioration in comparison to the results obtained by the other participating systems. The top two performing systems per task are shown in the table. In relation to wall-time execution, our system obtained the best results in all the tasks in which we participated. In addition, our aprroach also obtained the best results considering metrics related to energy consumption and the amount of emissions.

Table 5
Top two performing systems per task in terms of efficiency-related metrics (Sort criterion: Duration Mean). The best results are shown in bold.

Task	Team	Duration Mean	Emissions Mean	Energy Consumed Mean	Cpu Count	GPU Count
Task 1.A	NLP-UNED	**0.6128**	**1.62E-06**	**8.54E-06**	16	1
	Xabi IXA	2.0440	6.51E-06	3.43E-05	40	4
Task 1.B	NLP-UNED	**0.6128**	**1.62E-06**	**8.54E-06**	16	1
	Xabi IXA	2.0440	6.51E-06	3.43E-05	40	4
Task 2.A	NLP-UNED	**0.7316**	**1.64E-06**	**8.65E-06**	16	1
	CIMAT-NLP-GTO	1.5420	1.20E-04	2.47E-04	80	8
Task 2.B	NLP-UNED	**0.7316**	**1.64E-06**	**8.65E-06**	16	1
	CIMAT-NLP-GTO	1.5420	1.20E-04	2.47E-04	80	8
Task 2.C	NLP-UNED	**0.7316**	**1.64E-06**	**8.65E-06**	16	1
	plncmm	4.2677	1.25E-05	3.44E-05	12	1
Task 3.A	NLP-UNED	**0.5775**	**1.31E-06**	**6.90E-06**	16	1
	CIMAT-NLP-GTO	1.6076	1.25E-04	2.56E-04	80	8
Task 3.B	NLP-UNED	**0.5775**	**1.31E-06**	**6.90E-06**	16	1
	CIMAT-NLP-GTO	1.6076	1.25E-04	2.56E-04	80	8

6. Conclusions and Future Work

This paper presents a model based on dataset relabeling through ANNs to address the task of early detection of different disorders in Telegram messages written in Spanish. The model is a variant of the one already proposed for the study of Reddit messages in English [12]. This time, we use as encoder a Multilingual model based on Transformer [17]. The proposed pipeline has achieved average results, highlighting the results obtained for the classification tasks related to depression (Task 2.C; second place) and anxiety (Task 3.A; third place). In addition, considering the results provided by the organizers, our model obtained the best execution times in all the tasks where it was tested and quite positive emission results (Table 5). Finally, except for Task 1.A, the relabeling system was useful for classification tasks, although it was hampered by the small size of the corpus and by the decision taken regarding the use of disjoint training and validation sets.

As future work, and given that the system has shown difficulties processing Task 1.A, a qualitative study of the error cases detected will be proposed in order to solve possible inconsistencies in the scoring and tagging functions. On the hand, other kind of multilingual encoders will be explored in order to reduce the impact on the execution time, especially when processing large datasets. Finally, we will try to explore in depth the set of parameters used for the construction of the navigable index.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32 and OBSER-MENH Project (MCIN/AEI/10.13039/501100011033 and NextGenerationEU/PRTR) under Grant TED2021-130398B-C21 as well as the project RAICES (IMIENS 2022).

References

[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2022: Early Risk Prediction on the Internet (Extended Overview), in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 821–850. URL: https://ceur-ws.org/Vol-3180/paper-66.pdf.

[2] A. Tsakalidis, J. Chim, I. M. Bilal, A. Zirikly, D. Atzil-Slonim, F. Nanni, P. Resnik, M. Gaur, K. Roy, B. Inkster, J. Leintz, M. Liakata, Overview of the CLPsych 2022 Shared Task: Capturing Moments of Change in Longitudinal User Posts, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, Seattle, USA, 2022, pp. 184–198. URL: https://aclanthology.org/2022.clpsych-1.16. doi:10.18653/v1/2022.clpsych-1.16.

[3] P. López-Úbeda, F. M. P. del Arco, M. C. Díaz-Galiano, L. A. U. López, M. T. M. Valdivia, Detecting Anorexia in Spanish Tweets, in: R. Mitkov, G. Angelova (Eds.), Proceedings of

the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, Varna, Bulgaria, September 2-4, 2019, INCOMA Ltd., 2019, pp. 655–663. URL: https://doi.org/10.26615/978-954-452-056-4_077. doi:10.26615/978-954-452-056-4_077.

[4] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of Depression-Related Posts in Reddit Social Media Forum, IEEE Access 7 (2019) 44883–44893. URL: https://doi.org/10.1109/ACCESS.2019.2909180. doi:10.1109/ACCESS.2019.2909180.

[5] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_50.pdf.

[6] A. M. Mármol-Romero, S. M. J. Zafra, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, A. Montejo-Ráez, SINAI at eRisk@CLEF 2022: Approaching Early Detection of Gambling and Eating Disorders with Natural Language Processing, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 961–971. URL: https://ceur-ws.org/Vol-3180/paper-76.pdf.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A Robustly Optimized BERT Pretraining Approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[9] S. Serrano, N. A. Smith, Is Attention Interpretable?, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 2931–2951. URL: https://doi.org/10.18653/v1/p19-1282. doi:10.18653/v1/p19-1282.

[10] C. Meister, S. Lazov, I. Augenstein, R. Cotterell, Is Sparse Attention more Interpretable? (2021) 122–129. URL: https://aclanthology.org/2021.acl-short.17. doi:10.18653/v1/2021.acl-short.17.

[11] H. Amini, L. Kosseim, Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media, in: E. Métais, F. Meziane, H. Horacek, P. Cimiano (Eds.), Natural Language Processing and Information Systems - 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24-26, 2020, Proceedings, volume 12089 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 225–235. URL: https://doi.org/10.1007/978-3-030-51310-8_21. doi:10.1007/978-3-030-51310-8_21.

[12] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, UNED-NLP at eRisk 2022: Analyzing gambling disorders in Social Media using Approximate Nearest Neighbors, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 -

Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 894–904. URL: https://ceur-ws.org/Vol-3180/paper-71.pdf.

[13] H. Fabregat Marcos, A. Cejudo, J. Martinez-romo, A. Perez, L. Araujo, N. Lebea, M. Oronoz, A. Casillas, Approximate Nearest Neighbour Extraction Techniques and Neural Networks for Suicide Risk Prediction in the CLPsych 2022 Shared Task, in: Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, Seattle, USA, 2022, pp. 199–204. URL: https://aclanthology.org/2022.clpsych-1.17. doi:10.18653/v1/2022.clpsych-1.17.

[14] 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).

[15] V. Schmidt, K. Goyal, A. Joshi, B. Feld, L. Conell, N. Laskaris, D. Blank, J. Wilson, S. Friedler, S. Luccioni, Codecarbon: Estimate and track carbon emissions from machine learning computing, https://github.com/mlco2/codecarbon, 2021.

[16] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal Sentence Encoder, CoRR abs/1803.11175 (2018). URL: http://arxiv.org/abs/1803.11175. arXiv:1803.11175.

[17] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. Hernandez Abrego, S. Yuan, C. Tar, Y.-h. Sung, B. Strope, R. Kurzweil, Multilingual Universal Sentence Encoder for Semantic Retrieval (2020) 87–94. URL: https://aclanthology.org/2020.acl-demos.12. doi:10.18653/v1/2020.acl-demos.12.

[18] M. Aumüller, E. Bernhardsson, A. Faithfull, ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms, Information Systems 87 (2020) 101374. URL: https://www.sciencedirect.com/science/article/pii/S0306437918303685. doi:https://doi.org/10.1016/j.is.2019.02.006.

[19] M. Aumüller, E. Bernhardsson, A. J. Faithfull, ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms, in: C. Beecks, F. Borutta, P. Kröger, T. Seidl (Eds.), Similarity Search and Applications - 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings, volume 10609 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 34–49. URL: https://doi.org/10.1007/978-3-319-68474-1_3. doi:10.1007/978-3-319-68474-1_3.

[20] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview), in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 864–887. URL: https://ceur-ws.org/Vol-2936/paper-72.pdf.