A Time-Aware Approach to Early Detection of Anorexia: UNSL at eRisk 2024

Horacio Thompson[1,2], Marcelo Errecalde[1]

[1]Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, C.P. 5700, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), San Luis, Argentina

Abstract

The eRisk laboratory aims to address issues related to early risk detection on the Web. In this year's edition, three tasks were proposed, where Task 2 was about early detection of signs of anorexia. Early risk detection is a problem where precision and speed are two crucial objectives. Our research group solved Task 2 by defining a *CPI+DMC approach*, addressing both objectives independently, and a *time-aware approach*, where precision and speed are considered a combined single-objective. We implemented the last approach by explicitly integrating time during the learning process, considering the $ERDE_\theta$ metric as the training objective. It also allowed us to incorporate temporal metrics to validate and select the optimal models. We achieved outstanding results for the $ERDE_{50}$ metric and ranking-based metrics, demonstrating consistency in solving ERD problems.

Keywords

Early Risk Detection, Transformers, Decision Policy, Mental Health

1. Introduction

According to the World Health Organization, approximately one in every eight people worldwide suffers from a mental disorder, with anxiety, depression, bipolar disorder, and eating disorders being the most prevalent [1]. Multiple studies have underscored the correlation between social media usage and mental health disorders [2, 3, 4, 5]. Early Risk Detection (ERD) on the Web consists of correctly identifying users who show signs of mental health conditions as soon as possible. The Early Risk Prediction on the Internet (eRisk) laboratory has addressed challenges related to ERD problems in its different editions. This year, three tasks were proposed [6, 7], Task 2 being about early detection of signs of anorexia.

ERD incorporates a significant complexity to standard classification problems since the users are analyzed post-by-post rather than processing the complete history. In recent editions of eRisk, our research group proposed solutions based on a *CPI+DMC approach*, considering ERD as a multi-objective problem. The goal is to find an optimal balance between correctly identifying at-risk users and minimizing the time needed to make reliable decisions, addressing precision and speed independently. It involves defining two components: one dedicated to solving a classification problem with partial information (CPI) and the other to deciding the moment of classification (DMC). Using this approach, we achieved interesting results in the 2021 [8], 2022 [9], and 2023 [10] eRisk editions. In the last two editions, we applied the BERT model [11] with an extended vocabulary for the CPI component and a decision policy based on the model's prediction history during user evaluation (*historic rule*) for the DMC component. Similarly, in [12], we applied these methods in MentalRiskES 2023, the first ERD challenge for the Spanish language. In this case, we used the BETO model [13], a variant of BERT for Spanish, and adjusted the *historic rule* according to the tasks to be solved, obtaining excellent results.

Although precision is crucial for ERD problems, as time progresses and decisions are delayed, speed becomes as important as precision. In this context, it could be proper to consider ERD as a combined single-objective problem, simultaneously considering precision and speed within the learning process. Proposals such as [8, 9] applied the EARLIEST architecture for time series [14] in ERD, aiming to balance both objectives through reinforcement learning. Since the advent of transformers [15], numerous studies have focused on enhancing user classification through the standard fine-tuning process, but

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
hjthompson@unsl.edu.ar (H. Thompson); merreca@unsl.edu.ar (M. Errecalde)

often without addressing speed as a crucial aspect in decision-making. Models are trained and validated differently than how they are evaluated in ERD challenges, making it difficult to select the optimal model. Therefore, we propose a *time-aware approach*, which involves applying the fine-tuning process considering the user information and time progress. By incorporating time into the learning process, it is possible to define a scenario similar to that used in the evaluation stage in ERD. This allows us to select the optimal model using temporal metrics such as ERDE_θ while avoiding the necessity of defining complex decision policies for an early detection environment.

Considering both approaches, we addressed Task 2 of the eRisk 2024 edition using three models, two of which implemented the *CPI+DMC approach*, while the third adopted the *time-aware approach*. We achieved the second-best score for the ERDE_{50} metric and the first-best place in several rankings-based metrics, as well as an acceptable performance compared to the mean among all teams for the F_1 and $F_{latency}$ metrics. Additionally, we were the second-fastest team to solve the proposed task. Therefore, our proposals demonstrated consistency when evaluated for the early detection of signs of anorexia.

2. Time-aware approach

We propose a method that tunes transformer-based models by incorporating time during the learning process. The process involves modifying the inputs by adding time, applying the fine-tuning process by defining ERDE_θ as the training objective, and then evaluating the selected best model in an ERD environment.

Data preprocessing. Time is explicitly included in the input samples, considering the number of posts that have been read until the moment (*delay*). For an arbitrary sample $\langle$input, label$\rangle$, the new input is defined as input $_{delay}$ = [CLS] input [TIME] $delay$ [SEP]. The [TIME] token is added to the model architecture, separating the post content of the moment it was read. For example, *I don't feel like eating* evaluated at $delay = 10$, the model would receive [CLS] *I don't feel like eating* [TIME] 10 [SEP].

Time-aware fine-tuning. The process is carried out in epochs with the typical training and validation stages. We propose to incorporate an evaluation scheme similar to the testing environment for ERD problems in both stages, where time is measured based on *delays*. At each *delay*, post windows of length M are evaluated by concatenating the current post with the previous M-1 posts. The *delays* are configured based on the window size. For instance, with M=10, *delay*=10 assesses posts 0 to 9, *delay*=20 assesses posts 10 to 19, and so forth until all users have been analyzed. The *loss* function is computed at the end of each *delay*, evaluating the post window for users still undergoing analysis, and then the gradient is propagated throughout the transformer. We used the ERDE_θ metric [16] to design a linear and differentiable *loss* function. This involved incorporating classification performance (CrossEntropy) and heavily penalizing delayed true positives based on a threshold θ. Thus, as the *loss* is minimized, ERDE_θ is also reduced, establishing it as the training objective. The validation stage follows a similar *delay* scheme, concluding each epoch with the calculation of *loss*, *accuracy*, and ERDE_θ metrics. These metrics can be weighted to select the optimal model for an ERD problem.

Model testing in ERD. The best model obtained can be evaluated using a *mock-server* tool[1], which simulates ERD environments through rounds of posts and answers submissions and calculates the results using different metrics. A client application was defined to interact with the server: when it receives a round of posts, the system preprocesses them by adding time, invokes the predictive model, and returns a response. We used a sliding post window, configured as in the learning stage, and a simple decision policy (*simple rule*): if the probability exceeds a *threshold* (the prediction limit probability for a positive user), a user at-risk alarm is issued; otherwise, the analysis should continue.

[1]Available at: https://github.com/jmloyola/erisk_mock_server.

3. Task resolution

Task 2 was conducted following the guidelines from previous editions. It consisted of two stages: a *training stage*, where participants experimented with the data provided by the organizers, followed by a *testing stage*. During the latter, an early environment was established, wherein each participant deployed a client application to interact with a server, retrieving user posts one by one and submitting responses using the proposed predictive models. In this section, we provide details of the datasets involved in the task, the models proposed by our team, and the results achieved.

3.1. Datasets

Table 1 shows the details of the corpora available to solve the task. The organizers used the eRisk2024 corpus to evaluate the participating models, while the other corpora were released for participants to train their models. In our case, we trained the models using the eRisk2019 and eRisk2018_train corpora, reserving the eRisk2018_test to evaluate them in an early environment. All datasets have an imbalance in the classes, with a similar proportion of positive users among them. In particular, the eRisk2024 corpus contains 784 users, with 12% being positive, and compared to eRisk2018_test, it has 40% more samples, a greater number of posts per user, and more words per post.

Table 1
Details of the corpora used to solve Task 2. The number of users (total, positives, and negatives) and the number of posts in each corpus are reported. The mean, minimum, and maximum number of posts per user and words per post in each corpus are detailed.

Corpus	#users			#posts per user			#words per post		
	Total	Pos	Neg	Mean	Min	Max	Mean	Min	Max
eRisk2024	784	92	692	626	10	2,001	35	1	19,668
eRisk2019	815	73	742	700	10	2,000	32	1	8,953
eRisk2018_train	152	20	132	558	9	1,999	34	1	7,404
eRisk2018_test	320	41	279	527	9	1,999	33	1	8,318

3.2. Models

We presented three models to address Task 2, of which two applied the *CPI+DMC approach*, and a third applied the *time-aware approach*. Training and validation were performed by splitting the data in an 80/20 ratio. Several preprocessing steps were applied to the data before training, such as converting texts to lowercase, transforming Unicode and HTML codes into their corresponding symbols, replacing websites with the 'weblink' token, and eliminating repeated words, among other steps. Subsequently, the best models were evaluated in an early environment using the *mock-server* tool. Below are the details of each proposal.

UNSL#0 - CPI+DMC approach
CPI component. Classic fine-tuning using the BERT model. Hyperparameters: Architecture='BERT-based-uncased', optimizer='AdamW', LR=3E-5, scheduler='LinearSchedulerWarmup', batch_size=8, and n_epochs=5. These values were chosen based on the models' performance according to the F_1 metric on the positive class.
DMC component. Decision policy based on the *historic rule*: if the current prediction and last M predictions exceed a *threshold*, the client application must issue a risky user alarm; otherwise, the analysis should continue. In addition, the rule has the *min_delay* parameter, which defines the moment when it will begin to apply. Hyperparameters: *threshold*=0.7, *min_delay*=10, M=10. These values were obtained considering temporal metrics ($ERDE_{50}$ and $F_{latency}$) when evaluating in an early environment.

UNSL#1 - CPI+DMC approach
CPI component. Classic fine-tuning using the BERT model with extended vocabulary. We applied the

SS3 model [17] to expand the BERT vocabulary, considering the most relevant words for the anorexia domain. We evaluated different models by varying the number of words, ultimately selecting the top 50 according to the confidence value of SS3 on the positive class. Thus, previously unknown words to BERT, such as *calories*, *binge*, *tulpa*, *purging*, *vegan*, *underweight*, and *overweight*, were included. Additionally, we used the same hyperparameters as UNSL#0, configuring LR=5E-5.
DMC component. Decision policy based on the *historic rule*, with the same hyperparameters as UNSL#0.

UNSL#2 - Time-aware approach
We applied the *time-aware* fine-tuning process by modifying the input texts and used $ERDE_{50}$ as the training objective since participants are usually evaluated using this metric. To configure *delays* during the training and validation stages, we used a post window with a size of 10 and truncated the maximum length of user history to 200 posts to prevent bias in the model's learning. The best model was selected by weighting the accuracy and $ERDE_{50}$ metrics, obtaining the following hyperparameters: Architecture='BERT-based-uncased', optimizer='AdamW', LR=3E-5, scheduler='LinearSchedulerWarmup', batch_size=8, and n_epochs=10. Furthermore, because the model learns the decision policy during training, we used a *simple rule* configured with *threshold*=0.7 and *min_delay*=10.

3.3. Results

In this section, we present the results obtained during the current edition of eRisk. A total of 44 proposals from 10 teams were submitted to solve Task 2 (Table 2). Our team completed the task in 7 hours by processing all user posts through the three models previously described. This performance positioned us as the second-fastest team to complete the task.

Table 2
Total time spent by each team for Task 2. The team name, number of models, and number of user posts processed are shown.

Team	#models	#posts processed	Total time
UMUTeam	5	2001	6h:34m
UNSL	3	2001	7h
BioNLP-IISERB	5	10	9h:39m
NLP-UNED	5	2001	9h:40m
ELiRF-UPV	4	2001	12h:27m
Riewe-Perla	5	2001	2 days + 11h:25m
GVIS	5	352	3 days + 12h:36m
SINAI	5	2001	3 days + 23h:49m
APB-UC3M	2	2001	6 days 21h:34m
COS-470-Team-2	5	1	-

Table 3 displays the outcomes achieved by our team according to decision-based metrics. The three models attained the second-best $ERDE_{50}$, being surpassed by Riewe-Perla#0, while NLP-UNED#1 achieved top results in both F_1 and $F_{latency}$. Our models yielded satisfactory results, surpassing the mean of all teams for the F_1, $ERDE_{50}$, and $F_{latency}$ metrics. Moreover, UNSL#2 (*time-aware* model) exhibited the same performance as UNSL#0 and UNSL#1 (*CPI+DMC* models) in the $ERDE_{50}$ metric. This fact underscores UNSL#2's ability to optimize its $ERDE_{50}$ during training and apply a simple decision policy to solve the task, in contrast to the other models obtained through conventional fine-tuning, followed by a more complex policy.

Considering the ranking-based metrics (Table 4), our team achieved the best results in multiple categories across different post counts (1, 100, 500, and 1000). The results were comparable to NLP-UNED#1 and outperformed Riewe-Perla#0 in all metrics. The only teams that showed acceptable results under these metrics were UNSL, NLP-UNED, and Riewe-Perla. In contrast, the overall performance of the other teams was considerably lower, as evidenced by the mean and median values. Additionally, UNSL#1 and UNSL#0 achieved better performance than UNSL#2.

Table 3
Decision-based evaluation results for Task 2. The best teams taking into account the F_1, $ERDE_5$, $ERDE_{50}$, and $F_{latency}$ are shown, as well as the *mean* and *median* values of the results report for CLEF eRisk 2024. Values in bold and underlined depict 1st and 2nd performance achieved in the challenge, respectively.

Model	P	R	F_1	$ERDE_5 \downarrow$	$ERDE_{50} \downarrow$	latencyTP$\downarrow$	speed	$F_{latency}$
UNSL#0	0.35	0.99	0.52	0.14	<u>0.03</u>	12	0.96	0.49
UNSL#1	0.42	0.96	0.59	0.14	<u>0.03</u>	12	0.96	0.56
UNSL#2	0.42	0.97	0.59	0.14	<u>0.03</u>	12	0.96	0.56
Riewe-Perla#0	0.45	0.97	0.62	**0.07**	**0.02**	6	0.98	0.6
NLP-UNED#1	0.67	0.97	**0.79**	0.09	0.04	14	0.95	**0.75**
Mean	0.34	0.78	0.42	0.12	0.07	8.50	0.84	0.41
Median	0.41	0.97	0.49	0.11	0.06	6.00	0.96	0.47

Table 4
Ranking-based evaluation results for Task 2. Results are reported according to the three classification metrics obtained after processing 1, 100, 500, and 1000 posts, respectively. Values in bold represent the best performance achieved in the challenge.

	1 post			100 posts			500 posts			1000 posts		
Model	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
UNSL#0	0.9	0.81	0.63	**1**	**1**	0.81	**1**	**1**	0.77	**1**	**1**	0.76
UNSL#1	**1**	**1**	**0.69**	**1**	**1**	0.8	0.9	0.81	0.69	0.8	0.88	0.72
UNSL#2	0.4	0.38	0.42	0.9	0.92	0.71	0.8	0.85	0.69	0.8	0.84	0.68
Riewe-Perla#0	0.5	0.47	0.17	0.7	0.62	0.74	0.7	0.62	0.74	0.7	0.62	0.75
NLP-UNED#1	**1**	**1**	0.44	**1**	**1**	0.89	**1**	**1**	**0.92**	**1**	**1**	**0.92**
Mean	0.31	0.29	0.2	0.33	0.31	0.31	0.27	0.26	0.26	0.27	0.27	0.27
Median	0.2	0.12	0.14	0.2	0.14	0.15	0	0	0.06	0	0	0.7

4. Conclusion

In this year's edition, we solved Task 2 related to the early detection of signs of anorexia. The models based on the *CPI+DMC approach* demonstrated robustness when evaluated in a new application domain. In turn, the *time-aware approach* allowed including the time during the learning process and optimizing the $ERDE_{50}$ metric, avoiding the necessity of employing more complex decision policies when evaluating the models in an early environment. This fact encourages us to delve deeper into this approach, as we believe further research is necessary on the combination of precision and speed as a single objective to address ERD problems.

References

[1] F. Charlson, M. van Ommeren, A. Flaxman, J. Cornett, H. Whiteford, S. Saxena, New who prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis, The Lancet 394 (2019) 240–248.

[2] F. Aliverdi, H. Farajidana, Z. M. Tourzani, L. Salehi, M. Qorbani, F. Mohamadi, Z. Mahmoodi, Social networks and internet emotional relationships on mental health and quality of life in students: structural equation modelling, BMC psychiatry 22 (2022) 1–10.

[3] P. K. Maulik, W. W. Eaton, C. P. Bradshaw, The effect of social networks and social support on common mental disorders following specific life events, Acta Psychiatrica Scandinavica 122 (2010) 118–128.

[4] J. Martínez-Líbano, N. González Campusano, J. I. Pereira Castillo, et al., Las redes sociales y su influencia en la salud mental de los estudiantes universitarios: Una revisión sistemática, REIDOCREA 11 (2022) 44–57.
[5] F. A. Nawaz, M. M. A. Riaz, A. Singh, Z. Arshad, H. Derby, M. A. Sultan, et al., Social media use among adolescents with eating disorders: a double-edged sword, Frontiers in Psychiatry 15 (2024) 1300182.
[6] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, Springer International, 2024.
[7] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, Grenoble, France, September 9th to 12th, 2024, CLEF 2024, CEUR Workshop Proceedings, 2024.
[8] J. M. Loyola, S. Burdisso, H. Thompson, L. C. Cagnina, M. Errecalde, Unsl at erisk 2021: A comparison of three early alert policies for early risk detection., in: CLEF (Working Notes), 2021, pp. 992–1021.
[9] J. M. Loyola, H. Thompson, S. Burdisso, M. Errecalde, Unsl at erisk 2022: Decision policies with history for early classification., in: CLEF (Working Notes), 2022, pp. 947–960.
[10] H. Thompson, L. Cagnina, M. Errecalde, Strategies to harness the transformers' potential: Unsl at erisk 2023, in: CLEF (Working Notes), 2023, pp. 791–804.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[12] H. Thompson, M. Errecalde, Early detection of depression and eating disorders in spanish: Unsl at mentalriskes 2023, in: IberLEF (Working Notes), 2023.
[13] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.
[14] T. Hartvigsen, C. Sen, X. Kong, E. Rundensteiner, Adaptive-halting policy network for early classification, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 101–110.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[16] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.
[17] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.