NLP-UNED at eRisk 2024: Approximate Nearest Neighbors with Encoding Refinement for Early Detecting Signs of Anorexia

Notebook for the eRisk Lab at CLEF 2024

Hermenegildo Fabregat[1,3], Daniel Deniz[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, Marqués de Valdeiglesias, 3, Madrid 28004, Spain

Abstract

This paper describes our participation in Task 2 (Early Detection of Signs of Anorexia) from the CLEF 2024 eRisk Workshop, addressed to detecting early signs of anorexia in Social Media users through the analysis of their posts. A relabelling step based on Approximate Nearest Neighbors (ANN) is performed for generating a training dataset annotated at message level instead of user level, and then contrastive learning techniques are applied for refining the previously generated vector representations of the messages. ANNs are used also for classification purposes, combined with the use of rules and heuristics focused on expanding the number of considered messages from the user for making the final decision. Our system obtains the best results in both the decision-based evaluation, with 9 percentage points over the second best system in terms of latency-weighted F1, and in the ranking-based evaluation, with the best scores for 11 out of the 12 metrics employed.

Keywords

Early risk detection, Anorexia, Approximate Nearest Neighbors, Contrastive Learning,

1. Introduction

In recent years, the analysis of social media for early detection of health risks has become an intriguing and significant area of research. Within this research field, the eRisk workshop, part of the Conference and Labs of the Evaluation Forum (CLEF) since 2017, has played a pivotal role. This workshop fosters collaborative efforts to develop innovative methodologies and practical solutions for the early identification of various health concerns, including eating disorders, self-harm, pathological gambling and depression, through the analysis of textual content on social media platforms. By analyzing social media posts and messages, researchers can obtain valuable insights to identify individuals at risk.

This paper details our approach to tackling Task 2 of the eRisk 2024 Workshop [1, 2]: Early Detection of Signs of Anorexia. In this task, systems must sequentially process messages posted by different users in Reddit forums, searching for early traces of anorexia, this is, detecting as soon as possible whether a user is at risk of suffering from anorexia. The task is a continuation of Task 2 of the eRisk 2018 Workshop [3] and Task 1 of the eRisk 2019 Workshop [4].

Building upon our previous work in the detection of pathological gambling [5, 6, 7], we have refined our system by incorporating contrastive learning techniques for fine-tuning the encoded representations of text messages written by the analyzed users. Additional heuristics have been also included in the system in order to expand the context of the user's messages, this way taking into account a larger

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

gildo.fabregat@lsi.uned.es (H. Fabregat); daniel.deniz@avature.es (D. Deniz); 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); 0000-0002-0313-2127 (D. Deniz); 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)

number of previous messages when making the final decision on whether the user is at risk. These improvements have proven to enhance the system's accuracy and reliability in detecting potential cases of anorexia from social media content.

The rest of the paper is structured as follows: Section 2 gathers information about previous research works related to early detection of risks, as well as systems participating in previous eRisk competitions. A brief description of the addressed task, and the dataset and evaluation metrics involved is presented in Section 3. The different components of the proposed system are described in Section 4, and the results obtained by this system are shown and analyzed in Section 5. Finally, Section 6 depicts some conclusions about the work, together with possible future lines of work regarding this research.

2. Related Work

The automatic detection of mental health issues is currently a hot research topic within machine learning, specifically regarding natural language processing. The availability of information sources with large amounts of data, such as social media, is enabling the development of new systems aimed at the early detection of these types of issues. Within this context, different evaluation frameworks and campaigns such as CLEF's eRisk [8], CLPSych [9] or IberLEF's MentalRiskES [10, 11] represent a significant effort by the scientific community to support the development and dissemination of these types of systems.

Anorexia nervosa (AN) is a severe eating disorder characterized by an inability to maintain a healthy body weight, often falling below 85% of the ideal weight. Individuals with AN obsess over weight gain, perceive their bodies as larger than they are, and engage in behaviors to sustain weight loss. This illness profoundly affects both mind and body, with sufferers placing significant importance on their shape and weight, intertwining their self-esteem with their body image [12]. The 2018 and 2019 CLEF eRisk competitions addressed the automatic detection of signs of anorexia in Social Media posts, encouraging the participating systems to develop techniques for determining whether a user can be classified as at risk of suffering from this illness. Although the stage of development of neural models was nowhere near the current level when the last edition of this task was held (2019), some of the best participating systems at that time used such models for their predictions. An ensemble approach with different neural attention-based models is used in [13] for feature extraction, and then combined with Support Vector Machines to determine the final decision. Deep learning models are also used in [14] for developing a time series dataset representing the evolution of the user's mood through time. Then, Bayesian inference is employed for performing the final classification. Other approaches obtained good results in the competition by using more classic machine learning methods such as statistical word-based techniques [15], or Support Vector Machines with customized feature sets based on emotions derived from the text [16] or content-based features from phrases with personal pronouns [17]. In general, and also based on the results obtained by our own participations in early risk detection tasks, systems not relying on deep learning techniques or large language models are also able to achieve good results [7].

Contrastive learning techniques can be defined as methods aimed to learn and refine effective representations of data by pulling semantically close neighbors together and pushing dissimilar ones apart [18]. One of the most important characteristics of contrastive learning is that the model learns by comparison, this is, it is not necessary for the instances whose representations are to be refined to be accompanied by their corresponding labels. Instead, these approaches only need to define the similarity distribution. This way, the model should learn to map together similar instances, while separating dissimilar instances in the embedding space [19]. These techniques have been successfully applied to computer vision problems [20] and natural language processing tasks [21], as well as to other domains such as audio or reinforcement learning [22]. Considering our system presented in previous eRisk competitions, based on approximate nearest neighbors with vector representations of text messages, exploring these techniques seems like a logical step for its improvement.

3. Task 2: Early Detection of Signs of Anorexia

As previously mentioned, we have participated in task 2 of the eRisk 2024 competition, denoted "Early Detection of Signs of Anorexia". In this task, participants have access to a training dataset containing the whole history of writings (Reddit posts) for a set of users. These users are annotated depending on whether they have explicitly mentioned to have been diagnosed with anorexia (positive users) or not (negative or control users). In the test stage, systems are asked to determine, as soon as possible, whether a new user is at risk of suffering from anorexia according to the user's writing history. In particular, for each new message of a user, systems must determine whether the user is positive or negative. Once a user is labelled as positive, the decision is considered to be final, and hence all subsequent labels assigned to this user are ignored. Systems must also assign, after each message, a score measuring the user's risk of suffering from anorexia. This score is considered for evaluation purposes even after a user has been labelled as positive.

The statistics of the test dataset used for evaluating systems participating in this task are shown in Table 1:

Table 1
Main statistics of test collection for task 2: Early detection of signs of anorexia.

	Anorexia	**Control**
Num. subjects	92	692
Num. submissions (posts & comments)	28,043	338,843
Avg num. of submissions per subject	304.8	489.6
Avg num. of days from first to last submission	$\approx$ 482	$\approx$ 971
Avg num. words per submission	28.5	21.4

System evaluation is conducted using two different paradigms: decision-based evaluation and ranking-based evaluation. Complete information about the employed metrics can be found in [23].

- Decision-based evaluation: This type of evaluation only attends to the label assigned by the system to each user (positive or negative), as well as the delay in determining that a positive user is indeed at risk of suffering from anorexia. For this aim, standard metrics used for classification such as precision, recall and F-Measure are combined with metrics that take into account this delay information. The early risk detection error metric ERDE [24] is also used, although their values have low interpretability. To overcome this, other metrics regarding the latency and speed on detecting true positives are also proposed, and a final *latency-weighted F1* measure is computed by weighting the F-Measure with these delay-related metrics.
- Ranking-based evaluation: The score assigned to each user by the system, after analyzing each received message, is used in this evaluation for computing ranking-based metrics. This is, users are ranked after K messages according to this score, and then standard ranking metrics such as $P@K$ and $NDCG@K$ are applied for measuring the performance of the systems.

Finally, the lapse of time employed by the system for processing the whole test dataset is also measured and reported, in order to illustrate the efficiency of the proposed systems.

4. Proposed System

The system developed for performing early detection of signs of anorexia is presented in this Section. In particular, the different components that constitute the complete system pipeline are enumerated and described in detail. The main differences with the original research, based on dataset relabelling and approximate nearest neighbors techniques, presented in [5], are the use of a contrastive learning technique for fine-tuning the embedding representations of the user's messages (Section 4.3), as well as the development of a set of heuristics for considering previous messages for the final classification, instead of only taking into account the last message received (Section 4.4).

4.1. Data representation

The encoder used in this work for obtaining embeddings representing each of the messages of a particular user is the Universal Sentence Encoder [25]. Through its use, all messages in the training dataset are transformed into 512-dimensional embeddings. The specific model used in the encoding is based on a Deep Average Network (DAN) [26], trained on different sources of data written in English, and normally used for generating vector representations of texts longer than words, i.e., sentences, phrases or short paragraphs.

4.2. Relabelling process

The relabelling process has been described in previous works [5, 7]. Its main objective is to generate a training dataset labelled at message-level, starting from the user-level annotation provided by the organizers. The intuition behind this decision, already tested in previous eRisk competitions devoted to detecting pathological gambling, is that message-level annotations can help the system to emit accurate alerts about the risk of a user of suffering from anorexia by analyzing the user's individual messages.

In this stage a technique for generating indexes based on approximate nearest neighbors (ANN) is applied, this way creating a data structure that allows us to obtain the N most similar messages to a specific one. Two different ANN approaches have been explored in this work: first, Annoy [27] is a partitioning method based on the use of hyperplanes that recursively divide the search space with random direction. The generated index has the shape of a binary tree, and through its use the most similar elements to a query can be easily retrieved. On the other hand, the Hierarchical Navigable Small World (HNSW) method, implemented by the Non-Metric Space Library (NMSLIB) [28] is a graph-based ANN technique. In this case, the search index has the form of a proximity graph in which nodes correspond to particular instances (in our case, messages), and edges define the neighborhood relationship. The main idea behind the use of this technique is that a neighbor's neighbor is likely to also be a neighbor of a particular instance. Nearest neighbor retrieval is then performed by using a best-first search strategy on the graph.

Once that the selected index has been built on all the messages composing the training dataset, we are able to retrieve all the desired nearest neighbors given a particular message. In the first iteration of the relabelling process, all messages are labelled as belonging to the same class (positive or negative) as the user that created them. Then, for each positive message M in the training dataset, a set of its K nearest neighbors is retrieved from the index. The message will be relabelled as negative only if at least J of those K nearest messages belong to the negative class. In our implementation, only positive messages can be relabelled as negative. This is due to the fact that only positive users can have negative messages, because if negative users had any positive message they would have been labelled as positive. Only messages containing title information, this is, messages representing the opening of a Reddit thread, are taken into account for generating our training dataset. This filtering allows us to focus on discussions originally initiated by the analyzed user, which are more likely to contain information about particular worries or calls for help from the user. Moreover, this also reduces the computational complexity of the system, while the final results do not significantly differ from those obtained by using the complete set of messages. The relabelling step is iteratively repeated until convergence is reached, this is, no new relabellings are done during an iteration. A random sample of 33% of the users in the original training dataset is employed for validation purposes, allowing us to explore the optimal values of the K and J parameters. Through this validation step, these values have been set to $K = 10$ and $J = 6$.

4.3. Contrastive Learning

After completing the relabelling process, we propose an additional technique in the encoding step of our system based on fine-tuning the generated embeddings representing the different messages. This fine-tuning relies on a contrastive learning technique [29], a method employed for maximizing the distance between embeddings of messages belonging to different classes and minimizing it when

the messages belong to the same class. In particular, in our system this is achieved by retraining the Universal Sentence Encoder used for generating the initial representations of the messages. However, during this retraining, we employ a particular type of loss function, known as triplet loss [30]. For each message in the training dataset, either labelled as positive or negative, a triplet (a, p, n) is created, being a the original message, p a message belonging to the same class, and n a message belonging to the opposite class. The triplet loss function used in our retraining is $\mathcal{L} = max(d(a, p) - d(a, n) + \alpha, 0)$, where d is a function measuring the distance between the generated embeddings. The distance function employed for this work is cosine distance. This implies that the main aim of the training process will be to minimize the distance between messages belonging to the same class and maximize the distance between messages belonging to different classes. An additional parameter α is included into the loss function in order to determine the minimum desired distance between positive and negative instances, considering a as reference instance.

The main idea behind the contrastive learning process is illustrated in Figure 1.

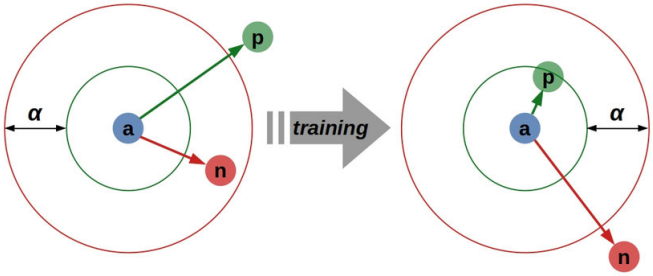

Figure 1: Contrastive learning with triplet loss: training is oriented to maximizing the distance between same class (p) and opposite class (n) instances with respect to a given anchor instance (a).

The different hyperparameters employed in the contrastive learning process are the following:

- **Number of instances:** 20 triplets (a, p, n) are generated for each message a, by randomly selecting positive (same class) and negative (opposite class) instances.
- **Batch size:** Batch size value is 32.
- **Learning rate:** The learning rate is set to $1e^{-5}$.
- **Epochs:** The number of epochs is 4.
- **Steps per epoch:** The number of steps per epoch is 128.
- **Margin:** The triplet loss margin (α) is set to 0.15 (normalized values are used for distances and margin).

With this configuration, a maximum of 128*32 (steps per epoch times batch size) instances are fed to the the network in each epoch. This implies that 128*32*4 = 16,384 instances are used for training. Hence, given the size of the training dataset, only a fraction of the generated triplets are effectively used for training. Also, not all instances are seen the same number of times.

4.4. Final classification

Once that the representation of the text messages is refined using contrastive learning techniques, the final classification step is somehow similar to the relabelling process described in Section 4.2. However, some additional heuristics have been added to this stage in order to consider more than one individual message for determining whether a user is at risk of suffering from anorexia.

Two new K and J parameters are calculated in this step for performing the final classification. Each time a new message M is received, the K nearest neighbors are retrieved. If at least J of those K neighbors are positive, the message, and hence the user, is directly classified as positive. Through the use of the validation split aforementioned, the values of these parameters have been set to $K = 19$ and $J = 19$ for the classification step.

As previously mentioned, we are also interested in analyzing whether the history of previous messages from the user can be useful for performing a more accurate classification. With this purpose, we have explored in more depth how assigning risk scores to the user after analyzing each message can affect the final classification. Besides the classification of the user as positive or negative, and regarding the ranking-based evaluation, a score is expected to be assigned to the user after receiving each message, representing the user's risk of suffering from anorexia. In our system, this score is computed by calculating the average distance between a received message M and all its nearest neighbors labelled as positive, $val = \frac{1}{k}\sum_{x=1}^{k} distance(U_x, M)$, where U_x is a message within the set of K nearest neighbors that is labelled as positive. The distance function employed returns values between 0 and 2, and hence the scoring assigned to the user is $score = (2 - val)$. This way, a message really close to its positive neighbors would receive a distance value of $val \approx 0$ and hence its score would be $score \approx 2$. This score is calculated for test messages classified as positive, but also for those classified as negative, and a buffer containing the scores of the N previous messages from the user is stored. The buffer is originally filled with zeros. Hence, if the system initially classifies a message as negative, the average score value for the last N messages is calculated, and the message (and user) will be classified as positive if this average is over a particular threshold S. The optimal values of N and S (this is, the message window considered and the score threshold) are also determined using the validation split and vary depending on the submitted run (see Section 5.1).

5. Results and Discussion

Main results achieved by the proposed system are presented in this Section. Experiments using the validation split are first depicted in order to justify the configurations selected for the submitted runs. Only decision-based evaluation, and more particularly, latency-weighted F1 values, were taken into account for tuning the hyperparameters through the validation split. Then, results obtained on the test dataset by the 5 different configurations selected are shown.

5.1. Validation and Selected Runs

As previously mentioned, a random split of 33% of the users in the training dataset is employed for validation purposes. Through these experiments we have confirmed that the use of the contrastive learning technique is able to improve all previous results obtained when using the Universal Sentence Encoder with no modifications for generating the embeddings. In particular, the latency-weighted F1 value of the best performing configuration that uses the original encoder is around 6% lower than the best performing system in our validation process. For this reason, we decided to use the contrastive learning encoder in all the submitted runs. In general, applying the relabelling method also improves the results with respect to not using it (this is, labelling all messages from a positive user as positive and all messages from a negative user as negative). However, we included a run that does not perform any relabelling in the test configurations, in order to compare results. The remaining parameters (values K and J in either relabelling or classification, and values N and S) have been adjusted by selecting the best performing configurations in the validation phase. As already stated, values of $K = 10$ and $J = 6$ during relabelling and $K = 19$ and $J = 19$ during classification showed the best results in this stage.

Table 2 shows the configurations of the proposed system, for each of the five runs allowed to be submitted in the task.

Column "ANN system" indicates the technique employed for building the nearest neighbor index: Annoy or NMSLIB. The type of encoder employed is always the one that refines the Universal Sentence Encoding with contrastive learning (CL_USE). Column "Relabel" indicates whether the relabelling step has been followed or not, while column "Heuristics" shows values for parameters N (window size) and S (decision threshold) in case the rules described in Section 4.4 have been employed, and "None" otherwise. It can be noticed how the best value for parameter S is always set to 1.0, this is, half the maximum scoring value that the average score for the N last messages can reach. Finally, we can

Table 2
Validation results: Configurations selected for the test phase.

Run	ANN system	Encoder	Relabel	Heuristics	Latency-weighted F1 (validation)
R0	Annoy	CL_USE	YES	$N = 7, S = 1.0$	0.6967
R1	Annoy	CL_USE	YES	None	0.6862
R2	Annoy	CL_USE	YES	$N = 5, S = 1.0$	0.6863
R3	Annoy	CL_USE	NO	None	0.6506
R4	NMSLIB	CL_USE	YES	$N = 7, S = 1.0$	0.6915

observe how the latency-weighted F1 metric is quite similar in this validation for all the proposed configurations, except for R3, which does not include the relabelling step.

5.2. Test results

The following tables illustrate the main results achieved by our system regarding the two types of evaluations considered, as well as the comparison with the other teams participating in the task. In particular, Table 3 shows results according to the decision-based evaluation.

Table 3
Test results: Results of the decision-based evaluation for task T2. Bold indicates the best result for each considered metric.

Team	Run	P	R	F1	ERDE5	ERDE50	Latency TP	Speed	Latency-weighted F1
NLP-UNED	0	0.64	0.97	0.77	0.09	0.04	13.00	0.95	0.73 (1)
NLP-UNED	1	0.67	0.97	**0.79**	0.09	0.04	14.00	0.95	**0.75 (1)**
NLP-UNED	2	0.63	0.97	0.76	0.09	0.04	12.00	0.96	0.73 (1)
NLP-UNED	3	0.63	0.98	0.77	0.09	0.03	11.00	0.96	0.74 (1)
NLP-UNED	4	0.63	0.97	0.76	0.09	0.04	14.00	0.95	0.72 (1)
BioNLP-IISERB	4	**0.73**	0.62	0.67	0.08	0.05	4.00	0.99	0.66 (2)
Riewe-Perla	0	0.45	0.97	0.62	**0.07**	**0.02**	6.00	0.98	0.60 (3)
ELiRF-UPV	0	0.43	0.99	0.60	0.10	0.04	12.00	0.96	0.57 (4)
UNSL	2	0.42	0.97	0.59	0.14	0.03	12.00	0.96	0.56 (5)
SINAI	0	0.21	0.92	0.34	0.10	0.07	3.00	0.99	0.34 (6)
APB-UC3M	0	0.17	0.99	0.28	0.15	0.08	9.00	0.97	0.28 (7)
UMUTeam	1	0.15	0.99	0.26	0.19	0.09	27.00	0.90	0.24 (8)
GVIS	1	0.12	**1.00**	0.22	0.12	0.10	**1.00**	**1.00**	0.22 (9)
COS-470-Team-2	0	0.00	0.00	0.00	0.12	0.12			(10)

As the table shows, all the configurations proposed for our system are able to overcome all participating systems in terms of latency-weighted F1. In particular, our best performing run, R1, is 9% ahead of the second best performing team. Although some other teams obtain slightly better results regarding precision and recall, the F1 and latency-weighted F1 values show that our proposal is the most robust across the considered metrics. Our system also obtains good results for some of the early risk detection metrics. In particular, it achieves the third best ERDE5 and second best ERDE50 values, although the latency and speed values are somewhat worse. It is particularly noticeable how all the proposed runs are able to obtain good results. This probably indicates that the main improvement proposed, which is the use of a contrastive learning technique for refining the embeddings representing text messages, has a powerful impact on the performance of our system. On the other hand, the use of heuristics for increasing the amount of information considered before classifying a message, does not seem to have that much impact on the final results. However, in the validation stage we have stated that when contrastive learning is not performed on the original embeddings, the use of these heuristics does positively influence the results. Therefore, future efforts should be focused on improving these rules.

Table 4 shows the main results on the ranking-based evaluation.

Once again, our system ranks first in this type of evaluation, for almost all the considered metrics, and for any of the proposed configurations. In particular, we are able to achieve perfect scores for $P@10$ and $NDCG@10$ after receiving 1, 100, 500 and 1000 messages, and the best results for $NDCG@100$

Table 4
Test results: Results of the ranking-based evaluation for task T2. Bold indicates the best result for each considered metric.

		1 writing			100 writings			500 writings			1000 writings		
		P@10	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**	**P@10**	**NDCG@10**	**NDCG@100**
NLP-UNED	R0	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
NLP-UNED	R1	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**	**0.92**
NLP-UNED	R2	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
NLP-UNED	R3	**1.00**	**1.00**	0.45	**1.00**	**1.00**	**0.91**	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
NLP-UNED	R4	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.91	**1.00**	**1.00**	0.91
UNSL	R1	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	0.80	0.90	0.81	0.69	0.80	0.88	0.72
Riewe-Perla	R0	0.50	0.47	0.17	0.70	0.62	0.74	0.70	0.62	0.74	0.70	0.62	0.75
GVIS	R1	0.40	0.37	0.40	0.30	0.32	0.42	0.00	0.00	0.00	0.00	0.00	0.00
ELiRF-UPV	R0	0.20	0.12	0.14	0.20	0.13	0.14	0.20	0.13	0.14	0.20	0.13	0.14
UMUTeam	R1	0.20	0.12	0.14	0.10	0.06	0.03	0.00	0.00	0.05	0.20	0.21	0.12
BioNLP-IISERB	R4	0.20	0.21	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ABP-UC3M	R0	0.00	0.00	0.03	0.40	0.56	0.26	0.00	0.00	0.09	0.00	0.00	0.13
SINAI	R3	0.00	0.00	0.07	0.10	0.07	0.06	0.00	0.00	0.07	0.00	0.00	0.07
COS-470-Team-2	R0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

after receiving 100, 500 and 1000. Only the UNSL team is able to beat our system for the $NDCG@100$ after seeing only the first message of each user. Together with our latency and speed values in the decision-based evaluation, this fact indicates that our system could be improved in terms of speed in finding true positives, this is, determining that a user is at risk of suffering from anorexia.

Finally, Table 5 shows some information regarding the number of runs submitted by the participating teams, the number of total writings processed by each team, and the total time employed in processing the messages.

Table 5
Participating teams, number of runs, number of user writings processed by the team, and lapse of time taken for the entire process.

Team	#Runs	#User Writings Processed	Lapse of Time (from 1st to last response)
BioNLP-IISERB	5	10	09:39
GVIS	5	352	3 days 12:36
Riewe-Perla	5	2001	2 days 11:25
UNSL	3	2001	07:00
UMUTeam	5	2001	06:34
COS-470-Team-2	5	1	-
ELiRF-UPV	4	2001	12:27
NLP-UNED	5	2001	09:40
SINAI	5	2001	3 days 23:49
APB-UC3M	2	2001	6 days 21:34

Compared to the other participating systems that processed the complete set of user writings, our system is the third best performing regarding execution times, the time interval being in the order of hours, in a similar manner to the best performing teams.

6. Conclusions and Future Work

This paper presents our participation in Task 2 of the CLEF eRisk 2024 competition: Early Detection of Signs of Anorexia. The developed system is a new version of the system designed for previous editions

of the competition, in which a relabelling method based on the use of approximate nearest neighbors (ANN) is applied on the training dataset, and the same ANN techniques are then used for classifying new messages and determining whether a user is at risk of suffering from a mental problem, in this case anorexia. The new improvements incorporated to the system is the use of contrastive learning techniques for fine-tuning the embeddings of the text messages, initially generated through a Universal Sentence Encoder, and the increasing of the amount of information employed for classification by including a set of rules or heuristics that consider a message window of N previous messages. The developed system is able to obtain the best results among the participating systems in terms of F-Measure and latency-weighted F1 (decision-based evaluation), as well as in terms of ranking-based evaluation metrics. In particular, all the tested configurations of the system overcome the second best participating team by around 9% of latency-weighted F1. In general, the main results indicate that the refinement of the vector representations obtained through contrastive learning techniques has been crucial for a better discrimination between positive and negative messages, thus leading the system to effectively determine when a message may indicate that the user is at risk of suffering from anorexia. On the other hand, expanding the message window considered for performing the final classification has not shown significant impact on the test results, although during the validation stage those configurations using these heuristics were able to obtain better overall results with respect to configurations only using one message for making a decision.

As mentioned in Section 5.1, future lines of work should focus on improving the rules designed for considering the history of messages before classifying a user. A trade-off must be found between the latency (this is, number of messages analyzed before emitting an alert) and the amount of information that should be gathered before making a decision. Also, the treatment of these previous messages can be improved: for instance, the current rules underestimate the weight of similar positive messages when few messages have been received, since the buffer of previous scores is initialized with zeros. This implies that even if a message is quite similar to positive messages its score is going to decrease when it is one of the first analyzed messages for a user. The current decision of selecting only the nearest positive messages for calculating the score can also be detrimental for the final results. More research should be done on the type of functions that better model the similarity of a given message with both positive and negative nearest neighbors, and its influence on the classification decision.

An additional future line of research involves further refinement of the embeddings used for representing users' messages. In particular, the hyperparameters used in the contrastive learning phase, described in Section 4.3 can be studied in greater depth through validation techniques, in order to search for optimal values. Additionally, different encoding models beyond the Universal Sentence Encoder could be also considered, exploring issues such as multilingualism or models that have already used contrastive learning techniques in their original training, like E5 [31].

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, OBSER-MENH Project (MCIN/AEI/10.13039 and NextGenerationEU"/PRTR) under Grant TED2021-130398B-C21 and EDHER-MED Project under grant PID2022-136522OB-C21, as well as by the Universidad Nacional de Educación a Distancia (UNED) within project SICAMESP (2023-VICE-0029).

References

[1] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet., Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024. Springer International Grenoble, France. (2024).

[2] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview)., 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).

[3] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: Early risk prediction on the internet (extended lab overview), in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/invited_paper_1.pdf.

[4] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at CLEF 2019: Early risk prediction on the internet (extended overview), in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_248.pdf.

[5] 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.

[6] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, NLP-UNED-2 at erisk 2023: Detecting pathological gambling in social media through dataset relabeling and neural networks, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 672–683. URL: https://ceur-ws.org/Vol-3497/paper-056.pdf.

[7] H. Fabregat, A. Duque, L. Araujo, J. Martínez-Romo, A re-labeling approach based on approximate nearest neighbors for identifying gambling disorders in social media, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 174–185. URL: https://doi.org/10.1007/978-3-031-42448-9_15. doi:10.1007/978-3-031-42448-9_15.

[8] J. Parapar, P. Martín Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet., Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece (2023).

[9] J. Chim, A. Tsakalidis, D. Gkoumas, D. Atzil-Slonim, Y. Ophir, A. Zirikly, P. Resnik, M. Liakata, Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts, in: A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, Y. Ophir (Eds.), Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 177–190. URL: https://aclanthology.org/2024.clpsych-1.15.

[10] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, L. A. U. López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2023: Early detection of mental disorders risk in spanish, Proces. del Leng. Natural 71 (2023) 329–350. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6564.

[11] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M. T. M. Valdivia, L. A. U. López, A. Montejo-Ráez, Mentalriskes: A new corpus for early detection of mental disorders in spanish, in: N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy, ELRA and ICCL, 2024, pp. 11204–11214. URL: https://aclanthology.org/2024.lrec-main.978.

[12] C. M. Bulik, L. Reba, A.-M. Siega-Riz, T. Reichborn-Kjennerud, Anorexia nervosa: definition,

epidemiology, and cycle of risk, International Journal of Eating Disorders 37 (2005) S2–S9.
[13] E. Mohammadi, H. Amini, L. Kosseim, Quick and (maybe not so) easy detection of anorexia in social media posts, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_74.pdf.
[14] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_126.pdf.
[15] S. G. Burdisso, M. Errecalde, M. Montes-y-Gómez, UNSL at erisk 2019: a unified approach for anorexia, self-harm and depression detection in social media, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_103.pdf.
[16] M. E. Aragón, A. P. López-Monroy, M. Montes-y-Gómez, INAOE-CIMAT at erisk 2019: Detecting signs of anorexia using fine-grained emotions, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_113.pdf.
[17] R. M. Ortega-Mendoza, D. I. H. Farías, M. Montes-y-Gómez, Ltl-inaoe's participation at erisk 2019: Detecting anorexia in social media through shared personal information, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_75.pdf.
[18] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, IEEE Computer Society, 2006, pp. 1735–1742. URL: https://doi.org/10.1109/CVPR.2006.100. doi:10.1109/CVPR.2006.100.
[19] P. H. Le-Khac, G. Healy, A. F. Smeaton, Contrastive representation learning: A framework and review, IEEE Access 8 (2020) 193907–193934. URL: https://doi.org/10.1109/ACCESS.2020.3031549. doi:10.1109/ACCESS.2020.3031549.
[20] T. Chen, S. Kornblith, M. Norouzi, G. E. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607. URL: http://proceedings.mlr.press/v119/chen20j.html.
[21] T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings, CoRR abs/2104.08821 (2021). URL: https://arxiv.org/abs/2104.08821. arXiv:2104.08821.
[22] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, CoRR abs/1807.03748 (2018). URL: http://arxiv.org/abs/1807.03748. arXiv:1807.03748.
[23] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 2021 2936 (2021) 864–887. URL: http://ceur-ws.org/Vol-2936/paper-72.pdf.
[24] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings, volume 9822 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 28–39. URL: https://doi.org/10.1007/978-3-319-44564-9_3. doi:10.1007/978-3-319-44564-9_3.
[25] 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.

[26] M. Iyyer, V. Manjunatha, J. Boyd-Graber, H. Daumé III, Deep unordered composition rivals syntactic methods for text classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1681–1691. URL: https://aclanthology.org/P15-1162. doi:10.3115/v1/P15-1162.

[27] E. Bernhardsson, Annoy: Approximate Nearest Neighbors in C++/Python, 2018. URL: https://pypi.org/project/annoy/, python package version 1.13.0.

[28] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, CoRR abs/1603.09320 (2016). URL: http://arxiv.org/abs/1603.09320. arXiv:1603.09320.

[29] N. Rethmeier, I. Augenstein, A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives, ACM Comput. Surv. 55 (2023) 203:1–203:17. URL: https://doi.org/10.1145/3561970. doi:10.1145/3561970.

[30] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada], 2005, pp. 1473–1480. URL: https://proceedings.neurips.cc/paper/2005/hash/a7f592cef8b130a6967a90617db5681b-Abstract.html.

[31] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, CoRR abs/2212.03533 (2022). URL: https://doi.org/10.48550/arXiv.2212.03533. doi:10.48550/ARXIV.2212.03533. arXiv:2212.03533.