OBSER-MENH at eRisk 2023: Deep Learning-Based Approaches for Symptom Detection in Depression and Early Identification of Pathological Gambling Indicators*

Notebook for the eRisk Lab at CLEF 2023

Juan Martinez-Romo[1,2,*], Lourdes Araujo[1,2], Xabier Larrayoz[3], Maite Oronoz[3] and Alicia Pérez[3]

[1]NLP & IR group (UNED). C/ Juan del Rosal, 16, 28040 Madrid (http://nlp.uned.es/)

[2]Instituto Mixto UNED-ISCIII (IMIENS)

[3]HiTZ Basque Center for Language Technologies - Ixa (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia (http://www.hitz.eus)

Abstract

Mental health problems, such as depression and pathological gambling, are conditions that can have very serious consequences if untreated, and cause the patient a lot of suffering. Research suggests that the way people write can reflect mental well-being and mental health risks, and social media provides a source of user-generated text to study. Early detection is crucial for mental health problems, and with this in mind the shared task eRisk was created. This paper describes the participation of the group OBSER-MENH on the T1 and T2 subtask at 2023. In the Task 1, participants had to provide rankings for the 21 symptoms of depression from the BDI-II Questionnaire and we used an approach based on the semantic textual similarity using Transformers. Task 2 consisted of sequentially processing pieces of evidence and detect early traces of pathological gambling as soon as possible. We implemented a penalty strategy in the loss function to deal with label imbalance. We combined three feed-forward neural networks with varying penalty values.

Keywords

early risk detection, depression detection, pathological gambling detection, natural language processing, semantic textual similarity

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

juaner@lsi.uned.es (J. Martinez-Romo); lurdes@lsi.uned.es (L. Araujo); xlarrayoz001@ikasle.ehu.eus (X. Larrayoz); maite.oronoz@ehu.eus (M. Oronoz); alicia.perez@ehu.eus (A. Pérez)

0000-0002-6905-7051 (J. Martinez-Romo); 0000-0002-7657-4794 (L. Araujo); 0000-0001-9097-6047 (M. Oronoz); 0000-0003-2638-9598 (A. Pérez)

1. Introduction

Mental health conditions, including depression and pathological gambling, have a significant impact on the lives of millions of individuals annually. Unfortunately, many individuals with these disorders fail to seek timely medical attention, resulting in unnecessary suffering. Some individuals are unaware of their need for treatment, while others avoid seeking help due to the associated stigma. Regardless of the reasons, untreated mental illnesses tend to worsen over time and can lead to severe consequences, such as substance abuse or even death.

Language serves as a fundamental means of communication between individuals, allowing for the transmission of intended messages while also conveying information about various aspects of oneself, such as upbringing, mood, and emotional well-being. Numerous studies have revealed a correlation between differences in language usage and writing style and the presence of mental health conditions [1, 2]. By employing Natural Language Processing (NLP) techniques, researchers can explore the application of language analysis to identify untreated mental health problems.

Social media platforms like Twitter and Reddit provide an extensive collection of user-generated texts, where individuals interact with friends, follow discussion groups, and express their thoughts and emotions. These platforms offer a vast amount of information that can be leveraged for NLP-based techniques with various purposes. Recent research has employed such techniques to automatically detect users who may be experiencing various mental health issues.

In the context of mental health, early intervention is particularly crucial as it enhances the likelihood of positive treatment outcomes. The longer a patient suffers without medical intervention, the higher the chances of experiencing associated risks. Early detection aids in identifying such cases before they escalate into more significant problems. While existing literature primarily focuses on detecting individuals who already have established mental health conditions, we contend that emphasizing early detection is vital for enabling prompt diagnosis and intervention.

To address this objective, the eRisk shared task was established. This shared task concentrates on the early detection of mental health problems within social networks. Previous editions of this initiative have targeted issues such as anorexia, self-harm, and pathological gambling. In the 2023 eRisk shared task [3], three subtasks were proposed, and this paper outlines our participation in subtasks T1 (Search for symptoms of depression) and T2 (Early Detection of Signs of Pathological Gambling).

The following sections are organized as follows: Section 2 describes the participation of our team in the task 1; Section 3 details our work for the participation in the task 2; Finally, Section 4 presents our conclusions and ideas for future work.

2. Task 1: Search for symptoms of depression

The objective of this task entails the arrangement of sentences extracted from a corpus of user-generated texts, based on their pertinence to a specific manifestation of depression. Participants will be requested to assign rankings to the 21 symptoms of depression as outlined in the Beck Depression Inventory (BDI-II) Questionnaire [4]. A sentence will be considered relevant to a

particular symptom if it contains information pertaining to the user's condition with respect to that symptom. In other words, a sentence may be deemed relevant even if it indicates that the user is experiencing no difficulties associated with the symptom in question.

2.1. Related Work

eRisk, under the umbrella of the international Conference and Labs of the Evaluation Forum (CLEF), is one of the most important initiatives towards early detection of mental health-related problems on the Internet. These evaluation campaigns provide a suitable environment for the automatic identification of early risks, the publication of corpora and collections, and the development of evaluation methodologies and metrics. In 2017 was proposed an initial pilot program [5] devoted to predicting depression, aiming to determine as soon as possible, whether an individual exhibits traces of depression through the analysis of posts in social media. In particular, the open-source platform Reddit was selected for building the collection of posts used in this task [6]. The task was also proposed in the 2018 and 2022 editions of eRisk. Different features have been considered for the detection of signs of depression in this task: emotion and sentiment words [7], readability features [8], words from depression-based lexicon or ontologies [9], linguistic metadata [10, 11] or information on the writing style [12], as well as features widely employed in NLP such as word embeddings or linguistic information like Part-of-Speech tagging. Deep learning models based on neural networks (Long-Short Term Memory or LSTMs and Convolutional Neural Networks or CNNs) have achieved the best results in this task [8, 10]. In the 2022 edition, many proposals were based on transformers, or combinations of these with other technologies [13].

The current edition differs from previous editions in that the focus is now on ranking of sentences from each user's writings according to their relevance with respect to a symptom of depression. We have resorted to the application of semantic similarity techniques to address it. The methods used for the calculation of semantic similarity include vector-based representations [14], graph-based models [15] and transformer-based models [16]. In this work we have preferred the transformer-based methods that currently provide the best results. These models use attention to capture interactions between words and generate high quality contextual representations. Similarity between texts can be computed using distance or cosine similarity between corresponding transformer representations. Specifically, we have used BERT (Bidirectional Encoder Representations from Transformers) [17], a language model that is known for its ability to capture the context and relationships between words. It uses a transformer architecture that applies multiple layers of attention and representational computations to process both the information before and after a given word. This allows BERT to capture the meaning and dependency of words in a broader context.

2.2. Dataset Description

The organizers provide to the participants a TREC formatted sentence-tagged dataset together with the BDI-II questionnaire.

The dataset is composed of a set of 3107 documents and each of these documents is composed of sentences. Below is an excerpt from document s_949.trec:

```
<DOC>
        <DOCNO>s_949_1409_1</DOCNO>
        <TEXT> I suspect I have depression.</TEXT>
</DOC>
```

2.3. Proposed Model

In order to calculate the similarity between the sentences and the 21 symptoms, we first transformed each piece of text into an embedding and then we calculated the cosine distance between the embeddings of each pair of pieces of text. The sentences are mapped such that symptoms with similar meanings are close in the vector space. For this, we use the Sentence Transformers (ST) framework [18] which employs a pre-trained BERT model to obtain the contextual representation of the sentences and symptoms and applies a mean pooling method to the output in such a way that it converts the embeddings of tokens to embeddings of sentences of a fixed size. This mean pooling technique, used by default in ST, is done by averaging the output embeddings.

We have used several models derived from BERT to obtain the embeddings that represent both the sentences and the text of every symptom in the questionnaire. BERT (and other transformer networks) output for each token in our input text an embedding. In order to create a fixed-sized sentence embedding out of this, the model applies mean pooling, i.e., the output embeddings for all tokens are averaged to yield a fixed-sized vector.

The models used in the different runs are as follows:

- all-mpnet-base-v2[1]: This model maps sentences and paragraphs to a 768 dimensional dense vector space.
- all-distilroberta-v1[2]: This model is pretrained from distilroberta-base model and fine-tuned on a 1B sentence pairs dataset.
- all-MiniLM-L12-v2[3]: This model is pretrained from the microsoft/MiniLM-L12-H384-uncased model and fine-tuned on a 1B sentence pairs dataset.

For models ”all-mpnet-base-v2” and ”all-distilroberta-v1”, two different runs were performed depending on the number of terms used to calculate the semantic similarity. Various sizes were used in the preliminary experiments and finally results were submitted for the first 90 and 20 terms respectively.

2.4. Results

In this section we analyze the task results of our participation. Once the runs from the participating teams have been submitted, organizers created the relevance judgements with the help of human assessors using pooling. They have used the resulting qrels to evaluate the systems with classical ranking metrics.

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[2]https://huggingface.co/sentence-transformers/all-distilroberta-v1
[3]https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

Table 1 shows the ranking-based evaluation for Task 1 using the majority voting approach. Based on the results obtained, the model with the best performance is "all-distilroberta-v1". In addition, across the different models, the use of the first 90 terms performed better than the use of 20.

Table 1
Ranking-based evaluation for Task 1 (majority voting) measured by Average Precision (AP), R-Precision, Precision at 10 and NDCG at 1000.

Team	Run	AP	R-PREC	P@10	NDCG@1000
OBSER-MENH	distilroberta-90-cos	**0.294**	**0.359**	**0.814**	**0.578**
OBSER-MENH	mpnet-90-cos	0.265	0.333	0.805	0.550
OBSER-MENH	mpnet-21-cos	0.120	0.207	0.471	0.365
OBSER-MENH	distilroberta-21-cos	0.158	0.249	0.543	0.418
OBSER-MENH	mini12-21-cos	0.114	0.184	0.305	0.329
Official Best results					
Formula-ML	SentenceTransformers_0.25	0.319	0.375	0.862	0.597

Table 2 shows the ranking-based evaluation for Task 1 using the unanimity approach. Based on the results obtained according this approach, the model with the best performance is also "all-distilroberta-v1". As in the previous evaluation method, the use of the first 90 terms performed better than the use of 20.

Table 2
Ranking-based evaluation for Task 1 (unanimity) measured by Average Precision (AP), R-Precision, Precision at 10 and NDCG at 1000.

Team	Run	AP	R-PREC	P@10	NDCG@1000
OBSER-MENH	distilroberta-90-cos	**0.281**	**0.344**	**0.652**	**0.604**
OBSER-MENH	mpnet-90-cos	0.252	0.337	0.643	0.575
OBSER-MENH	distilroberta-21-cos	0.135	0.216	0.390	0.413
OBSER-MENH	mini12-21-cos	0.099	0.165	0.214	0.329
OBSER-MENH	mpnet-21-cos	0.101	0.189	0.319	0.366
Official Best results					
Formula-ML	SentenceTransformers_0.1	0.293	0.350	0.686	0.611

3. Task 2: Early Detection of Signs of Pathological Gambling

The aim of this task is to identify signs of pathological gambling as earlier as possible [3]. With that aim, participants are given a set of social media posts that have to be processed in the order they were written. The sooner a pathological gambler is detected by the system, the better. The task is also addressed as a ranking decision problem, rather than assigning labels 0 (non pathological gambler) or 1 (pathological gambler), a score of the estimation of the risk to suffer such a disorder is computed.

3.1. Related Work

Early detection of signs of pathological gambling is crucial to increase the effectiveness of psychological therapies and be able to help patients at an early stage of the disease.

When this assignment was first introduced in 2021, eRisk did not offer labeled data [19]. Both last and this year the task has been carried out with provided labeled training data [20].

Regarding the methods employed to address this task, in 2021 UNSL team, the team attaining the best results, analyzed three different early alert policies based on standard classification models, rule-based algorithms and deep learning models. In this case, the best result was achieved with an SVM [21]. On the same year, UPV-Symanto team made use of BERT transformers in order to detect pathological gamblers [22].

Last year, with training data made available, UNSL team proposed a variant of the previous year incorporating two score normalization steps reducing the runtime and improving the model performance [23]. The BLUE [24] team, who achieved similar results, trained a BERT classifier with additional training data. The team that achieved the best F-score was NLP-UNED [25] with a 0.868 F1. They used an Approximate Nearest Neighbour approach in order to assign post level labels.

There was, as well, an approach based on FFNN, by the SINAI group [26], attaining an F-score of 0.8. They fed the FFNN with vectors that encapsulated emotions, semantic information, lexical diversity and volumetry of the posts. For what us regards we found that this approach was comprehensive and opted to explore it to get base-learners in order to build an ensemble model.

3.2. Dataset Description

The dataset is composed by a set of XML files. Each file comprises the user posts each with a separated title, and the timestamp in which they were posted. The training files consist of the test posts provided in the two previous editions: CLEF eRisk 2021 and 2022. Labels are given at user level being positive (denoted as "1") if the user is classified as a pathological gambler and negative ("0") otherwise. The test set, provided iteratively by a server, has the same source as the training data. Both the training and test data are quantitatively presented in Table 3.

Table 3
Quantitative description of the train and test data sets.

	Train		Test	
	Pathological Gamblers	*Control*	*Pathological Gamblers*	*Control*
Num. subjects	245	4,182	103	2,071
Num. total posts	2,298,412		1,102,871	

As presented in table 3, in both train and test sets the class distribution is unbalanced. Pathological gamblers do not exceed 6% and 5% in training and test sub-sets respectively.

The texts were pre-processed, as follows: first the title and posts were concatenated, next, stop-words were removed.

In order to obtain a numeric representation of the posts, we employed the Universal Sentence Encoder (USE) [27]. This encoder, on it's Dynamic Aggregation of Network (DAN) variant,

generates an embedding of dimension 512 as the output.

3.3. Proposed Model

Our approach focuses on tackling skewed class-distribution and also prevent over-fitting and entails an ensemble combining variants of a base-model. In what follows we describe the base models involved and next the combination strategies involved and the strategies incorporated to deal with class imbalance.

3.3.1. Base models

Our approach consists of a simple Feed Forward Neural Network (FFNN) implemented using Python torch library [28] fed by the posts represented by USE [27] encoder. Softmax activation function is employed to predict the class at post level. With respect to the practical details, we paid attention to over-fitting and tried to prevent it in two ways. On the one hand, AdamW optimizer was selected in the training process; on the other hand, 0.1 dropout was set.

The class distribution is clearly imbalanced, with 'Control' being the majority label (nearly 20 times more frequent than the target 'Pathological gamblers'), as shown in Table 3. As a consequence the network can become skewed, and consequently obtain low accuracy in the minority class. To address the class imbalance, we have applied a loss function based on cross-entropy also implemented in torch library.

During training, when labels are assigned at post level, first, the loss is calculated using cross entropy loss between the predicted post level labels (denoted as y'_k) and ground truth labels (denoted as y_k). Additionally, a user-level label is estimated from the estimated post-level labels. If any of the post level predicted labels, $(y'_{k1}, \dots, y'_{kn})$, is positive (being n the amount of posts for user k), the user is considered positive ($Y'_k = 1$), as in expression (1). This user-level label is then compared with the user's ground truth label to train the model.

$$Y'_k = \begin{cases} 1, & \text{if } \exists \text{ i in } 1 <= i <= n \quad y'_{ki} > 0.5 \\ 0, & otherwise \end{cases} \tag{1}$$

As an additional measure in order to address over-fitting, we implemented a penalty at user-level. That is, when a user is positive but was predicted as negative, the cross entropy loss is penalized by a scalar. The higher the penalty, the more will the loss be penalized. Our base penalty function is presented in (2). We consider all post level predictions of a user k as the vector $\vec{y'_k} = (y'_{k1}, \dots, y'_{kn})$ and ground truth post level labels as vector $\vec{y_k} = (y_{k1}, \dots, y_{kn})$, w is the penalty weight. The penalization is applied whenever a user label is predicted incorrectly ($Y'_k \neq Y_k$)

$$loss = \begin{cases} CrossEntropyLoss(\vec{y'_k}, \vec{y_k}) & \text{if } Y'_k = Y_k \\ CrossEntropyLoss(\vec{y'_k}, \vec{y_k}) \cdot w & \text{otherwise} \end{cases} \tag{2}$$

3.3.2. Ensemble models

In order to analyze the impact of different penalty values to deal with label imbalance, we implemented an ensemble model. We trained three base-models varying the values of the

penalty weight (w_0=5, $w_1 = 3$, $w_2 = 2$). Note that we get a post level prediction for a user post (y'_{kn}) by each of the three base-models involved, denoted as M_l (with $0 \leq l \leq 2$). These predictions are denoted as $M0_{y'_{kn}}$, $M1_{y'_{kn}}$, $M2_{y'_{kn}}$ respectively.

Next, we built three alternative ensemble models (M^e), each, yielding its prediction ($M^e_{y'_{kn}}$) using the following techniques:

- **Max voting:** The final prediction for a post is the maximum among the base-models involved as in (3).

$$M^{eMax}_{y'_{kn}} = \max_l Ml_{y'_{kn}} \tag{3}$$

- **Average:** The prediction represents the average of all models predictions, as in expression (4).

$$M^{eAvg}_{y'_{kn}} = \frac{1}{m} \sum_l Ml_{y'_{kn}} \tag{4}$$

- **F-score weighted average:** On this technique different weights are assigned to each model predictions based on their F-score (f_l) to promote each models contribution to the average as in (5).

$$M^{eWav}_{y'_{kn}} = \frac{\sum_l f_l \cdot Ml_{y'_{kn}}}{\sum_l f_l} \tag{5}$$

3.4. Results

In a preliminary experiment we tested the base-models with and without the penalty weight presented in (2), and also varying the penalty weight. Experimentally, we found that it was worth using the penalty weight and the best weighting values resulted to be w_0=5, $w_1 = 3$, $w_2 = 2$. Next we combined the selected base learners (each with the aforementioned penalty weight) and thus built ensemble models following the three alternative combination methods explored in section 3.3.2. Table 4 shows the configuration used on each run. The first two rows refer to base-models while the last three rows refer to ensemble models combined the base-learners mentioned in the corresponding column.

Table 4
Configurations for different runs. M_0, M_1 and M_2 refers to the generated models with different penalty values w_0=5, $w_1 = 3$ and $w_2 = 2$. The ensembles involved are: M^{eWav} as in (5); M^{eMax} as in (3); M^{eAvg} as in (4).

Run	Base-learners	Ensemble
OBSER-MENH 0	M_0	None
OBSER-MENH 1	M_1	None
OBSER-MENH 2	M_0, M_1, M_2	M^{eWav}
OBSER-MENH 3	M_0, M_1, M_2	M^{eMax}
OBSER-MENH 4	M_0, M_1, M_2	M^{eAvg}

Table 5 and table 6 show the results of each of the runs. If we look at decision based performance, table 5, despite the fact of employing different penalty values all the models'

behavior is similar. The Recall is 1 in all the runs and the precision is close to 0. This means that the model is effectively identifying the positive gamblers. Nevertheless, the low precision indicates a high number of false positives. In other words, the model predicts a user as positive in almost all cases.

Table 5
Decision based performance results measured by Precision, Recall, F1, $ERDE_5$, $ERDE_{50}$, latency, speed and latency-weigthed F1.

Team	P	R	F1	ERDE 5	ERDE 50	latency	speed	latency-weighted F1
OBSER-MENH 0	0.048	**1.000**	0.092	0.064	0.049	3.0	0.992	0.092
OBSER-MENH 1	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091
OBSER-MENH 2	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091
OBSER-MENH 3	0.048	**1.000**	0.092	0.063	0.049	3.0	0.992	0.091
OBSER-MENH 4	0.048	**1.000**	0.092	0.063	0.050	3.0	0.992	0.091

In table 6 we present the ranking based performance results. These results concern the users' level of risk estimated from the writings processed so far. As in previous results, our models show the same behaviour with a little improvement in runs 0,3 and 4 when taking into account 1000 writings. It achieves good results in P@10 and NDCG@10. However, when the ranking is calculated accross 100 sample users (NDCG@100) our results worsen.

Table 6
Ranking-based performance. It measures the system's efficiency after having classified 10, 100, 500 and 1000 writings.

Team	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
OBSER-MENH 0	1.00	**1.00**	0.64	**1.00**	**1.00**	0.55	**1.00**	**1.00**	0.48	0.90	0.94	0.50
OBSER-MENH 1	1.00	**1.00**	0.65	1.00	1.00	0.56	1.00	1.00	0.49	0.80	0.88	0.50
OBSER-MENH 2	1.00	**1.00**	0.65	1.00	**1.00**	0.56	1.00	1.00	0.49	0.80	0.88	0.50
OBSER-MENH 3	1.00	**1.00**	0.64	1.00	1.00	0.56	1.00	1.00	0.48	0.90	0.94	0.50
OBSER-MENH 4	1.00	**1.00**	0.65	1.00	**1.00**	0.56	1.00	1.00	0.49	0.90	0.94	0.50

4. Conclusions and Future Work

Our participation in the 2023 edition of eRisk has focused on two tasks. On the one hand, task 1 addressed the challenge of searching for symptoms of depression. On the other hand, task 2 dealt with the problem of early detection of signs of pathological gambling.

Regarding task 1, we tackle the issue as a semantic similarity task by leveraging transformers and employing pre-trained models to compute the similarity. Our approach has yielded a highly effective method, exhibiting optimal performance and being the second team in the ranking of best results. In the future we would like to address the problem of ambiguity in the use of certain terms that can confuse models based on semantic similarity.

Our proposed approach for task 2: early detection of signs of pathological gambling, was an ensemble between three models. The aim of the ensemble models was to analyze the different penalty weights applied on the FFNN. The implemented model was effective when predicting the risk level of the patients, obtaining proficient results in that task. Nevertheless, it wasn't able to detect users that don't have signs of pathological gambling.

Further research should be done to deal with the label imbalance and improve our method. It would be interesting to asses the effects of oversampling the minority class or to use Synthetic Minority Oversampling Technique (SMOTE) for artificial examples. Moreover, results suggest that to perform an ensemble on this scenario isn't beneficial, showing similar results in all cases. Therefore, a natural progression of this work is to make an ensemble of different Deep Learning methods that show differences on their individual performance.

Acknowledgments

OBSER-MENH, with subprojects GELP (TED2021-130398B-C21) and LOTU (TED2021-130398B-C22) are funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR. In addition, this work was partially funded by the Spanish Ministry of Science and Innovation (DOTT-HEALTH/PAT-MED PID2019-106942RB-C31, European Comission (FEDER) and INDICA-MED PID2019-106942RB-C32), by the Basque Government (IXA IT-1570-22, Ikasiker BOPV 11/07/2022); and by EXTEPA within Misiones Euskampus 2.0.

References

[1] M. De Choudhury, S. Counts, E. Horvitz, Social Media as a Measurement Tool of Depression in Populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 47–56.

[2] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological Aspects of Natural Language Use: Our Words, Our Selves, Annual Review of Psychology 54 (2003) 547–577.

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

[4] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of general psychiatry 4 (1961) 561–571.

[5] D. E. Losada, F. Crestani, J. Parapar, erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8, Springer, 2017, pp. 346–360.

[6] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings 7, Springer, 2016, pp. 28–39.

[7] I. Abdou Malam, M. Arziki, M. Nezar Bellazrak, F. Benamara, A. El Kaidi, B. Es-Saghir, Z. He, M. Housni, V. Moriceau, J. Mothe, et al., Irit at e-risk, CEUR Workshop Proceedings, 2017.
[8] M. Trotzek, S. Koitka, C. M. Friedrich, Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression., in: CLEF (Working Notes), 2017, p. 2017.
[9] F. Sadeque, D. Xu, S. Bethard, Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection, in: CEUR workshop proceedings, volume 1866, NIH Public Access, 2017.
[10] M. Trotzek, S. Koitka, C. M. Friedrich, Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia., in: CLEF (Working Notes), 2018.
[11] E. Campillo-Ageitos, J. Martinez-Romo, L. Araujo, Uned-med at erisk 2022: depression detection with tf-idf, linguistic features and embeddings, Working Notes of CLEF (2022) 5–8.
[12] F. Cacheda, D. F. Iglesias, F. J. Nóvoa, V. Carneiro, Analysis and experiments on early detection of depression., CLEF (Working Notes) 2125 (2018) 43.
[13] H. Srivastava, L. N. S, S. S, T. Basu, Nlp-iiserb@erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media., in: CLEF (Working Notes), 2022, p. 2022.
[14] H. T. Nguyen, P. H. Duong, E. Cambria, Learning short-text semantic similarity with word embeddings and external knowledge sources, Knowledge-Based Systems 182 (2019) 104842.
[15] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.
[16] A. Lauscher, I. Vulić, E. M. Ponti, A. Korhonen, G. Glavaš, Specializing unsupervised pretraining models for word-level semantic similarity, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1371–1383. URL: https://aclanthology.org/2020.coling-main.118. doi:10.18653/v1/2020.coling-main.118.
[17] 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).
[18] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[19] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2021: Pathological gambling, self-harm and depression challenges, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43, Springer, 2021, pp. 650–656.
[20] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings, Springer, 2022, pp. 233–256.
[21] 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.
[22] A. Basile, M. Chinea-Rios, A.-S. Uban, T. Müller, L. Rössler, S. Yenikent, M. A. Chulvi-Ferriols, P. Rosso, M. Franco-Salvador, Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet, in: Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021, CEUR, 2021, pp. 908–927.
[23] J. M. Loyola, H. Thompson, S. Burdisso, M. Errecalde, Unsl at erisk 2022: Decision policies with history for early classification (2022).
[24] A.-M. Bucur, A. Cosma, L. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, in: CEUR Workshop Proceedings, CEUR-WS, 2021. doi:10.13140/RG.2.2.25060.50567.
[25] 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: Conference and Labs of the Evaluation Forum, 2022.
[26] A. M. Mármol-Romero, S. M. Jiménez-Zafra, F. M. Plaza-Del-Arco, M. D. Molina-González, M.-T. Martín-Valdivia, A. Montejo-Ráez, Sinai at erisk@clef 2022: Approaching early detection of gambling and eating disorders with natural language processing, in: CEUR Workshop Proceedings, volume 3180, CEUR-WS, 2022, pp. 961–971.
[27] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, 2018. arXiv:1803.11175.
[28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.