UMUTeam at eRisk@CLEF 2023 Shared Task: Transformer Models for Early Detection of Pathological Gambling, Depression, and Eating Disorders

Ronghao Pan[1,*,†], José Antonio García-Díaz[2,†] and Rafael Valencia-García[1,†]

[123]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain

Abstract

This paper describes the participation of the UMUTeam in the eRisk shared task organized at CLEF 2023. We have addressed the three proposed tasks. For this purpose, several approaches have been used such as the fine-tuning of a sentence transformers model to measure the severity of the signs of eating disorders, the relevance search through the depression-domain lexicon and the sentence transformers models, the fine-tuning of the evaluated Large Language Models, and the development of an ensemble learning model for the detection of signs of pathological gambling. In Task 1, our team ranked in 8th place in majority voting Ranking-based evaluation and 5th place in unanimity Ranking-based evaluation. For Task 2, we have reached the 6th position in the decision-based evaluation ranking. As for Task 3, we have obtained 4th place, out of 4 participants, however in some evaluation metrics we are in the top 5 of all submissions such as RS and WCS.

Keywords

Depression Detection, Transformers, Large Language Model, Sentence Transformers, Natural Language Processing

1. Introduction

Social networks have become a useful source of data for the early detection of mental disorders and health problems due to a large amount of content posted on social media on a daily basis. Early detection technologies have applications in a variety of areas, but their most common use is in the field of health and safety [1]. Recent events such as the COVID-19 pandemic have highlighted the vulnerability of our mental health to confinement and uncertainty [2]. For example, in [3], the authors explored the relationship between COVID-19 and gambling. Another example is [4], in which the authors performed an aspect-based sentiment analysis of topics related to COVID-19 using an infectious disease ontology that includes concepts about depression and mental health.

State-of-the-art technologies related to Natural Language Processing (NLP) can generate early alerts when an adolescent's handwriting begins to show signs of depression or a psychological

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece
ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)
0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

disorder. eRisk (Early Risk Prediction On The Internet) is a task to explore techniques, evaluation methodology, effectiveness metrics, and practical applications for the early detection of mental health problems that manifest themselves in the way people write and communicate on the Internet, particularly in user-generated content. The eRisk@CLEF 2023 [5] focuses on the early detection of signs of depression, pathological gambling, and eating disorders. This shared task was defined using the test collection and evaluation metrics were proposed. In this edition, three tasks have been proposed:

- **Task 1: Search for symptoms of depression.** This is a novel task that consists of classifying sentences from a collection of user writings according to their relevance to a given symptom of depression. In this case, the 21 signs of depression from the BDI questionnaire are considered, where a sentence is considered relevant if it provides information about the user's state in relation to the symptom.
- **Task 2: Early Detection of Signs of Pathological Gambling.** It consists of sequentially processing evidence and detecting early signs of pathological gambling or gambling addiction as early as possible. This task focuses on texts written on social media. Therefore, texts are processed in the order in which they are created, so systems that effectively perform this task could be used to sequentially monitor user interactions on blogs, social networks, or other types of online media.
- **Task 3: Measuring the severity of the signs of Eating Disorders.** This task consists of estimating the level of characteristics associated with an eating disorder diagnosis from a thread of user submissions. To assess the range and severity of characteristics associated with an eating disorder diagnosis using the Eating Disorder Examination Questionnaire (EDE-Q). The EDE-Q is a 28-item self-report questionnaire adapted from the semi-structured interview Eating Disorder Examination (EDE).

This work presents the participation of the UMUTeam in all tasks. For this purpose, several approaches have been used such as the fine-tuning of a sentence transformers model to measure the severity of the signs of eating disorders, the relevance search through the depression-domain lexicon and the sentence transformers models, and the fine-tuning of the LLMs (Large Language Models) for the detection of signs of pathological gambling. The rest of the paper is organized as follows. Section 2 presents the task and the provided dataset. In addition, the methodology of our proposed system for addressing each task is described. Secondly, Section 3 shows the results obtained, and a discussion of them is presented. Finally, Section 4 concludes the paper with some conclusions and perspectives for future work.

2. Task description and Methodology

This section presents different eRisk tasks aimed at the early detection of symptoms of depression, signs of pathological gambling and eating disorders, as well as the dataset provided to perform them. In addition, the methodology of our proposed system for addressing each task is described.

2.1. Task 1: Search for symptoms of depression

This is a new task, not presented in the previous edition of eRisk, which consists of ranking sentences from a collection of written sentences according to their relevance to a depressive symptom. For this purpose, the BDI questionnaire consisting of 21 multiple-choice questions was used to measure the severity of depression. The data provided by the organizers is a dataset tagged with sentences in TREC format based on data from previous editions of eRisk. The dataset consists of a collection of 3,107 user writings from different languages with a total of 3,608,504 sentences. Despite the use of the *langdetect* library[1], a total of 18 languages have been identified in the dataset, of which English has the largest share with a total of 99.05%. Therefore, this task will be approached from two different perspectives: multilingual, considering all languages, and monolingual, using only the English language dataset.

2.1.1. System and methods

In order to address this task, we used a depression-domain lexicon approach to select texts most closely related to depression, and models of sentence transformers were used to rank the 21 BDI depression symptoms [6]. The pipeline used for this task is shown in Figure 1 and can be described as follows. First, the dataset is processed by removing all punctuation and replacing mentions with *@USER*, hashtags with *#HASHTAG*, and URL references with *URL*. Second, with more than 3 million sentences across all users, the following selection techniques were used to eliminate sentences unrelated to depression.

1. Stop words have been removed and a count of the number of mentions, hashtags, and URLs in the text has been made.
2. Texts with words from the lexicon defined by [7] have been selected.
3. Texts containing more than three useful words were selected, i.e., texts containing more than 3 words after removing stop words and excluding reference words (URLs, mentions, and hashtags).

After the text selection process, the dataset has been reduced to 296,985 sentences. Finally, sentence transformers models have been used to rank the 21 depression symptoms of the BDI. Sentence transformers models are deep learning models specially designed for sentence-level embedding and similarity calculations. These models are able to encode sentences into fixed-length vectors that capture their semantic meaning. The resulting embedding can be used to measure similarity or relatedness between phrases, enabling a wide range of downstream applications such as information retrieval, clustering, and question-answering systems. In this case, we have considered Task 1 as a question-answering task, where the questions are the questions of the BDI questionnaire linked to their possible answer, and the users' writing is considered as a possible answer to the question-answering system [8]. In this way, we can obtain the degree of relationship of each text with respect to the 21 symptoms of depression of the BDI.

[1]https://pypi.org/project/langdetect/

For this purpose, we have used two pre-trained models: *paraphrase-multilingual-MiniLM-L12-v2* [8] model for the multilingual approach and *all-MiniLM-L6-v2*[2] for the monolingual approach.

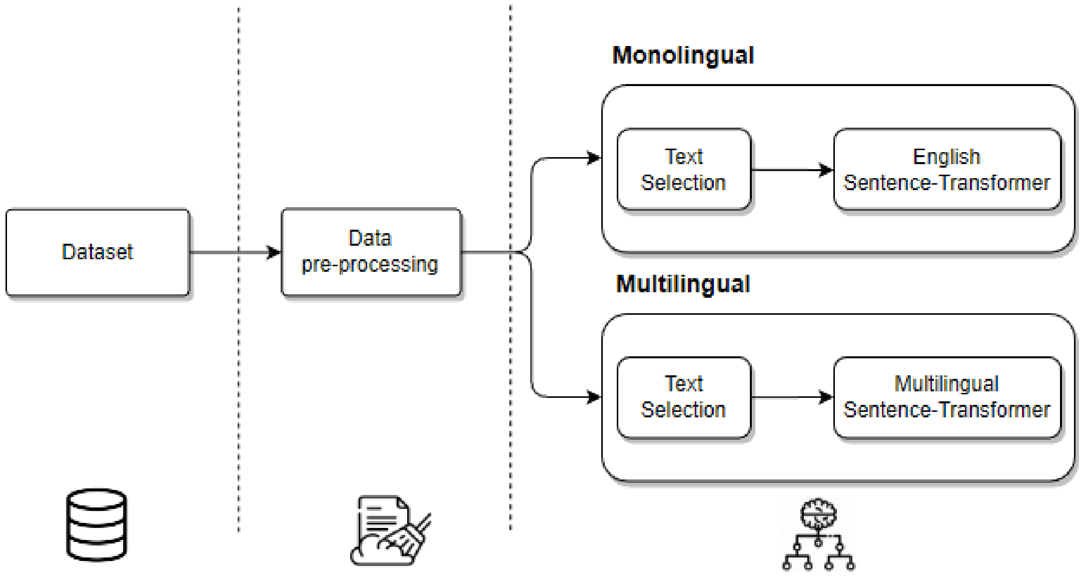

Figure 1: System architecture of Task 1.

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

This task is a continuation of the eRisk 2022's T1 task [9], which is the early detection of gambling risk by processing social media posts in strict order of publication. For the early detection evaluation, the participant system was tasked with sequentially reading, and processing messages from several users in order to generate a response to get the next posts. The training dataset is composed of data from the eRisk 2021 [10] and 2022 [9] editions. In the 2021 version, the dataset consists of 2,348 subjects, of which 54,674 posts by 164 subjects are categorized as positive for gambling addiction, and about 1M posts by 2,184 subjects are not categorized as addicted. In the 2022 version, the dataset contains a total of 2,079 subjects, of which 14,627 posts by 81 subjects are positive in gambling addiction, and about 1M posts by 1,998 subjects are categorized as negative. As can be seen, if we put the two datasets together there is a total of more than 2M messages among all subjects, so we have performed a selection process to reduce the training set. The selection techniques used are as follows:

1. All mentions, URL references, and hashtags have been removed.
2. Identify and remove sequences of type amp;format=png, amp;s=7b66887b445eb00d7d842b15e15f4759f3deb03d, etc.

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3. Texts containing more than three useful words were selected, i.e., texts containing more than 3 words after removing stop words and reference words (URLs, mentions, hashtags, and amp expressions).

To address this task, we followed a supervised learning approach. To train our model, we used the two datasets obtained after the selection process. It is worth mentioning that the organizers only provided training data, so we selected a custom split for validation. The customized validation split is created using stratified sampling, in order to keep the balance between labels. The final set of training and validation is shown in Table 1.

Table 1
The distribution of the training and validation split for task 2.

	Addicted	No Addicted	Total
Training	45 872	478 166	524 037
Validation	11 468	119 542	131 009

2.2.1. System and methods

This task is addressed as a binary decision problem, where users' messages are labeled as positive (label 1, i.e., addiction detected) or negative (label 0, no addiction detected). To address this task, we have followed a supervised learning approach, which is based on the fine-tuning of a pre-trained model such as BERT or RoBERTa to perform the classification task. The ensemble learning approach has also been evaluated by training the outputs of the trained RoBERTa-base [11] for pathological gambling classification with emotion features using Long Short-Term Memory Networks (LSTM). Emotion is a feature that focuses on measuring different emotions that are expressed in posts, such as fear, anger, joy, love, sadness, or surprise. It is worth noting that our research group has some experience dealing with emotion analysis in Spanish [12] but not in English. We obtain these emotions using a pre-trained language model based on the Transformer architecture called *distilbert-based-uncased-emotion*, which is a DistilBERT-based model fine-tuned on an annotated Twitter corpus on emotions [13]. For early detection, we have evaluated two strategies:

- **Conservative strategies**: It consists of making a decision after processing all the user's messages. In this case, the most repeated label will be taken.
- **"Courageous" strategies**: It consists of making a decision when the probability of signs of pathological gambling in a user's messages exceeds a certain threshold.

For our participation in eRisk Task 2, we have fine-tuned RoBERTa-base with 4 epochs, 16 of train batch size, 0.01 weight decay, and 2e-5 learning rate. For ensemble learning, once the RoBERTa base is trained, we create an LSTM model using the output of the classification model with the emotion features. The network architecture has four 1D convolutional layers assembled by 64 and 128 filters with an Exponential Linear Unit (ELU) as the activation function. In addition, an LSTM with 2 linear layers has been added after the convolutional layers, as shown in Figure 2.

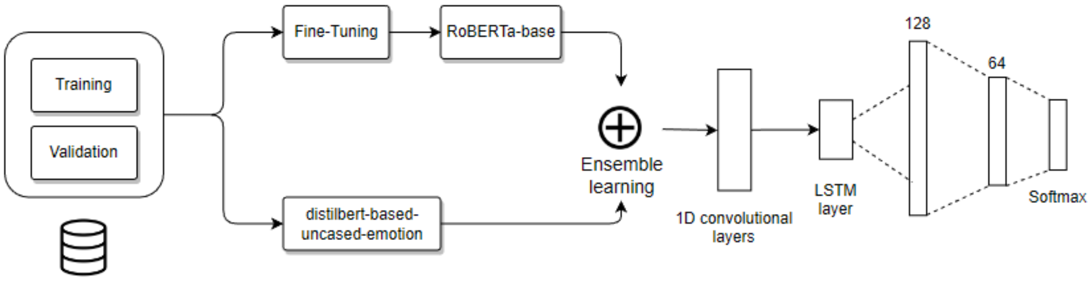

Figure 2: System architecture of Task 2.

2.3. Task 3: Measuring the severity of the signs of Eating Disorders

This is a new task compared to previous editions of eRisk, which consists of estimating the level of features associated with an eating disorder diagnosis from a thread of user submissions. The organizers provided a history of each user's posts and a standard eating disorder questionnaire. Thus, the main goal of this task is to predict the user's possible responses to the questionnaire from his or her posting history. The questionnaire is defined by the Eating Disorder Examination Questionnaire (EDE-Q), a 28-item self-report questionnaire adapted from the semi-structured interview Eating Disorder Examination (EDE). In this case, we need to predict the possible answers to questions 1-12 and 19-28. The dataset is composed of 28 history posting of different users and their answers to the EDE-Q questionnaire. For this task, we used a fine-tuning approach of a sentence transformers model based on textual similarity to estimate the similarity between the possible answers (user thread text) with each question of the EDE-Q. For this, we processed the user texts and associated them to the 22 questions with a score defined by the user's answers to the questionnaire. To obtain the scale-based score, we defined the following intervals for each possible answer of the questionnaire.

0. NO DAYS / not at all (0 to 0.1)
1. 1-5 DAYS / slightly (0.1 to 0.2)
2. 6-12 DAYS / slightly (0.2 to 0.3)
3. 13-15 DAYS / moderately (0.3 to 0.4)
4. 16-22 DAYS / moderately (0.4 to 0.5)
5. 23-27 DAYS / markedly (0.5 to 0.7)
6. EVERY DAY / markedly (0.7 to 1.0)

Therefore, in the training set, each text is associated to the questions and a score is defined which will be a random value in the corresponding interval according to the user's answer. In addition, we selected a customized split for validation with a ratio of 80-20. In the training set it has 121,932 text-question relations with their scores and in the validation set it has 30,484.

2.3.1. System and methods

The pipeline used to address this task can be described as follows. First, the dataset was processed by removing contractions, mentions, hashtags, URLs, and AMP expressions, and extracting emoji features using the Python *emoji* library. Second, we fine-tuned the *multi-qa-mpnet-base-dot-v1*[3] model with cosine similarity as the loss function, 3 epochs, and 100 warm-up steps. *multi-qa-mpnet-base-dot-v1* is based on the MPNet (Multilingual Pretrained BERT) architecture, which in turn is based on the BERT (Bidirectional Encoder Representations from Transformers) model.

3. Results

This section describes the systems submitted by our team in each run and shows the results obtained in each task.

3.1. Task 1: Search for symptoms of depression

We submitted two runs for this task. The runs have the same structure but differ in minor aspects of configuration.

- **Run 0**: This run consists of using a multilingual approach, i.e., using a pre-trained multilingual sentence transformers model (see Section 2.1) to rank symptoms of depression.
- **Run 1**: This run is similar to the previous one, but uses a pre-trained sentence transformers model with an English corpus (see Section 2.1) since 99% of the dataset is in English.

Based on the results reported by the organizers, we have extracted our scores, which are shown in Table 2, 3. The metrics used to evaluate this task are Average Precision (AP), R-Precision, Precision at 10, and NDCG at 1000.

Table 2
Ranking-based evaluation for Task 1 (majority voting).

Run	AP	R-PREC	P@10	NDCG@1000
0	0.073	0.140	0.495	0.222
1	0.054	0.122	0.362	0.191

Evaluation based on majority vote ranking, the AP obtained in our Run 0 is 0.073 and in Run 1 is 0.054. We can see that the multilingual approach has obtained better results than the monolingual approach. In this case, we have reached eighth with a total of 10 participants with 39 runs.

However, in the evaluation based on the unanimity ranking, we obtained the fifth-best result with a 0.059 AP in Run 0 and 0.044 in Run 1.

[3]https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

Table 3
Ranking-based evaluation for Task 1 (unanimity).

Run	AP	R-PREC	P@10	NDCG@1000
0	0.059	0.125	0.333	0.209
1	0.044	0.110	0.210	0.175

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

For task 2, we presented a total of 5 runs with different early detection strategies and classification models.

- **Run 0**: This run consists of running a classification model obtained through the fine-tuning RoBERTa-base within which the set of posts used has been preprocessed. The strategy used for early detection is "courageous" with a threshold of 0.8, i.e., the decision is made when it is identified that one of the user's posts has a probability greater than 0.8 of being addicted to gambling.
- **Run 1**: This run uses the same classification model as Run 0, but uses the conservative strategy for early detection, which involves making a decision after processing all user posts.
- **Run 2**: In this run, we used the same early detection strategy as in Run 0, but with an ensemble learning model that combines the output of the classification model (fine-tuned RoBERTa-base) with the emotion features.
- **Run 3**: This run uses the same classification model as Run 2, but uses the conservative strategy to make a decision about whether a user is addicted to gambling or not.
- **Run 4**: This run has the same structure as Run 2, but changing the brave strategy threshold to 0.5.

The results obtained with the approaches explored by our team are shown in Table 4 and 5.

Table 4
Results of UMUTeam for Task 2 in decision-based evaluation including the precision (P), recall (R), and F1-score (F1). Other metrics considering the performance of the methods are also included.

Run	*P*	*R*	*F1*	$ERDE_5$	$ERDE_{50}$	$latency_{tp}$	*speed*	$latency_w F1$
1	1.000	0.388	0.559	0.047	0.043	94.5	0.651	0.364
0	0.086	1.000	0.158	0.039	0.029	2.0	0.996	0.157
2	0.048	1.000	0.092	0.057	0.044	2.0	0.996	0.091
3	0.593	0.311	0.408	0.048	0.045	80.0	0.701	0.090
4	0.048	1.000	0.091	0.053	0.045	2.0	0.996	0.223

From the decision-based evaluation table (see Table 4), we can see that the classification models without ensemble learning perform better and that our conservative strategy (Run 1 and Run 3) has obtained a higher F1 than the "courageous" strategy (Run 2, 4 and 5). The best result was obtained with Run 1, with an F1 of 0.559, which we achieved to rank 6th with 11 participants and 49 submissions. In terms of P, R we also reached maximum values in several runs.

Table 5
Results of UMUTeam for Task 2 in ranking-based evaluation (only 1 writing result reported).

Run	P@10	NDCG@10	NDCG@100
0	1.00	1.00	0.63
1	0.00	0.00	0.03
2	0.00	0.00	0.00
3	0.00	0.00	0.03
4	0.00	0.00	0.00

With respect to ranking-based evaluation, Run 0 has achieved the best result with top value in P@10 and NDCG@10, and 0.63 in NDCG@100 (in 1 writing results). It can be noticed that these scores are lower for our system above 1 writing. We believe this is due to the number of submissions our system requires to trigger an alert, especially in the conservative strategy that needs to process all text to make a decision.

3.3. Task 3: Measuring the severity of the signs of Eating Disorders.

For this task, we have presented a run based on fine-tuning a pre-trained sentence transformers model with dataset processing and emoji feature extraction (see Section 2.3.1). The results of our team are shown in Table 6.

Table 6
Results of UMUTeam for Task 3 in performance results.

Run	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
0	2.194	0.800	2.027	2.288	1.777	2.412	2.556	2.135

For this task, a total of 4 teams participated with a total of 15 submissions, but none of them passed the baseline. In this case, the organizers have defined three baseline variants: all 0s and all 6s, which consists of sending the same answer (0 or 6) for all questions, and the average. When averaging all the scoring metrics, our team got the worst results. However, for metrics such as RS and WCS, we are in the top 5 of all submissions.

4. Conclusion

This paper describes the participation of the UMUTeam in Task 1, Task 2, and Task 3 of the eRisk@CLEF 2023 edition. In these editions, two new tasks have been proposed that focus on searching for symptoms of depression and measuring the severity of signs of eating disorders. The organizers have also proposed a continuation of Task 1 from last year's edition, which aims to detect signs of pathological gambling as early as possible. For Task 1, we have treated it as a question-answering problem, so we have proposed two approaches, one monolingual and one multilingual. Both approaches use a pre-trained sentence transformers model to measure the level of each depressive symptom in users' text collections. In this task, we have achieved 8th place in majority voting ranking-based evaluation and 5th place in unanimity ranking-based evaluation. For Task 2, we have developed a classification model by fine-tuning RoBERTa and an ensemble learning model that combines the output of a pre-trained transformer-based model for emotion feature extraction with the output of the gambling addiction classification model we have created (fine-tuned RoBERTa). In addition, we have evaluated two early detection strategies (conservative and courageous). As for the results for Task 2, we have reached the 6th position in the decision-based evaluation ranking. In terms of Task 3, we have created a question-answering model by fine-tuning a pre-trained sentence transformers model.

In future work, we plan to analyze other early detection techniques to improve the performance of Task 2 and other sentence transformers models to improve the results of Task 3. As discussed in Section 2.1.1, we have selected sentences that have more than 3 useful words as a method to reduce the dataset, so we would ignore the sentences "I am sad", etc. Therefore, as a future line we propose to test another threshold such as 1 or 2 useful words for the selection method. Moreover, we consider that it is relevant to consider the relationship between depression signs and hate-speech [14] and demographic and psychographic traits of the authors of the messages [15].

Acknowledgments

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033.

References

[1] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk: Early risk prediction on the internet, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 343–361.

[2] M. M. Hossain, S. Tasnim, A. Sultana, F. Faizah, H. Mazumder, L. Zou, E. L. J. McKyer, H. U. Ahmed, P. Ma, Epidemiology of mental health problems in covid-19: a review, F1000Research 9 (2020).

[3] M. Brodeur, S. Audette-Chapdelaine, A.-C. Savard, S. Kairouz, Gambling and the covid-19 pandemic: A scoping review, Progress in Neuro-Psychopharmacology and Biological Psychiatry 111 (2021) 110389.

[4] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 641–657.

[5] 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, Thessaloniki, Greece, Springer, 2023.
[6] M. von Glischinski, R. von Brachel, G. Hirschfeld, How depressed is "depressed"? a systematic review and diagnostic meta-analysis of optimal cut points for the beck depression inventory revised (bdi-ii), Quality of Life Research 28 (2019) 1111–1118.
[7] G. Li, B. Li, L. Huang, S. Hou, Automatic construction of a depression-domain lexicon based on microblogs: Text mining study, JMIR Medical Informatics 8 (2020).
[8] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[9] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, erisk 2022: pathological gambling, depression, and eating disorder challenges, in: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II, Springer, 2022, pp. 436–442.
[10] 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.
[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[12] J. A. Garcıa-Dıaz, R. Colomo-Palacios, R. Valencia-Garcıa, Umuteam at emoevales 2021: Emotion analysis for spanish based on explainable linguistic features and transformers, in: Proceedings of the Iberian Languages Evaluation Forum (Iber-LEF 2021), CEUR Workshop Proceedings, Málaga, Spain, 2021, pp. 59–71.
[13] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, Y.-S. Chen, Carer: Contextualized affect representations for emotion recognition, in: Conference on Empirical Methods in Natural Language Processing, 2018, pp. 3687–3697.
[14] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems (2022) 1–22.
[15] J. A. García-Díaz, R. Colomo-Palacios, R. Valencia-García, Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020, Future Generation Computer Systems 130 (2022) 59–74.