Transfer Learning for Automated Responses to the BDI Questionnaire

Christoforos Spartalis[1], George Drosatos[2] and Avi Arampatzis[1]

[1]Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece
[2]Institute for Language and Speech Processing, Athena Research Center, Xanthi 67100, Greece

Abstract

This paper describes the participation of the DUTH-ATHENA team of Democritus University of Thrace and Athena Research Center in the eRisk 2021 task, which focuses on measuring the level of depression based on Reddit users' posts. We address this task using both feature-based and fine-tuning strategies for applying BERT-based representations. In the feature-based approaches, we examine the possibilities of a SBERT model based on RoBERTa, pre-trained on Natural Language Inference (NLI) data and fine-tuned on STSb dataset to leverage transfer learning to depression-level estimation, and we achieve promising results. One of our runs ranks first in Average Hit Rate (AHR), while the others rank among the best four in the other evaluation metrics. Also, for the fine-tuning approach, we propose two predictive models that are built upon RoBERTa, which provide directions for future optimizations.

Keywords

Transfer Learning, SBERT, RoBERTa, Depression Level, Social Media, Reddit.

1. Introduction

In the last decade, the social interactions taking place in the digital world have increased [1]. This development expands the potential of monitoring systems that detect users who suffer from mental health conditions. Several studies have focused on this purpose using data from social media platforms, such as Facebook [2], Twitter [3], Reddit [4], and others. CLEF eRisk[1] contributes in this direction.

CLEF's eRisk lab [5] launched in 2017 introducing the test collection and evaluation metrics proposed in [6]. Since 2017, the eRisk shared tasks pave the way for early detection of signs of depression, self-harm, and anorexia [7]. Recently, a new challenge proposed concerning pathological gambling.

Since 2019, eRisk organizes a task oriented to automatically filling a depression questionnaire based on user interactions in social media. The Beck's Depression Inventory (BDI) questionnaire [8] consists of 21 questions which assess the presence of feelings and mental states, such as:

- Sadness, pessimism, agitation, irritability, guilty, and punishment feelings.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
chrispar@ee.duth.gr (C. Spartalis); gdrosato@athenarc.gr (G. Drosatos); avi@ee.duth.gr (A. Arampatzis)
https://www.drosatos.info (G. Drosatos); http://www.aviarampatzis.com (A. Arampatzis)
0000-0001-8228-4235 (C. Spartalis); 0000-0002-8130-5775 (G. Drosatos); 0000-0003-2415-4592 (A. Arampatzis)
© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1]CLEF eRisk – Early risk prediction on the Internet (https://erisk.irlab.org)

CEUR-WS.org/Vol-2936/paper-84.pdf

- Self-dislike, self-criticalness, worthlessness, tiredness, and indecisiveness.
- Changing in sleep patterns and appetite.
- Loss of pleasure and energy, loss of interest in sex and in general.
- Crying, failure, concentration difficulty, suicidal thoughts and wishes.

The performance of the approaches proposed in previous years to handle this task can be found in [9, 10]. Our approaches that are discussed in this paper are based on modifications of the BERT model [11]. We addressed the eRisk task as a downstream task and deployed both of the existing strategies for applying pre-trained language representations to it (i.e. feature-based and fine-tuning). Regarding the first one, we extract Reddit post representations from a SBERT pre-trained model [12] on the basis of which we build the predictive models. Regarding fine-tuning, we update the parameters of a RoBERTa pre-trained model [13] using the datasets provided by eRisk in previous years.

This paper is structured as follows. In Section 2, an overview of related works is provided. In Section 3, we describe the given eRisk datasets. In Section 4, we present our approaches to measuring the severity of depression signs. In Section 5, we present and discuss our scores in comparison with the best ones. Finally, in Section 6, we summarize our contributions and present some thoughts for future work.

2. Related Work

Some previous contributions [14, 15, 16, 17] to the eRisk shared tasks employed standard machine learning models, such as SS3 [18], topic modeling algorithms (LDA [19] and Anchor [20]), and neural models (Contextualizer [21], Deep Averaging Network [22], RNN [23], CNN [24, 25, 26], and BiLSTM [27, 28, 29]).

Several studies (e.g., [30]) suggest that pre-training a language model on a large corpus can provide widely applicable representations of words, which can be used in related tasks. These language models encode textual data into high dimensional vector representations, which are known as embeddings. In this way, the problem of lack or inadequacy of task-dedicated training data could be alleviated.

Some authors [31, 32] took advantage of these methods to automatically extract signals from social media activity concerning depression and anorexia. Some of them included pre-trained representations, extracted from GloVe [33], BERT [11] or Universal Sentence Encoder [34], as additional features to their task-specific architectures, whilst others [35, 36] fine-tuned OpenAI GPT [37] and XLM [38] pre-trained models.

3. Dataset

The task 3 of eRisk 2021 is a continuation of 2019's task 3 and 2020's task 2. The datasets of the two previous years are annotated and provided by the organizers upon request. In other words, the subjects' answers to the 21 questions of the BDI questionnaire are known. Thus, we utilized this data to evaluate our methods and select the best-performing ones for the eRisk 2021 challenge. Moreover, the approaches needed training or fine-tuning are solely based on eRisk 2019 and 2020 datasets. Table 1 shows the number of subjects and their posts per year. Their depression categories to which they belong vary from year to year, as shown in Figure 1.

Table 1
Makeup of the eRisk datasets.

	# of subjects	# of posts
eRisk 2019 (Task 3)	20	10,941
eRisk 2020 (Task 2)	70	35,562
eRisk 2021 (Task 3)	80	30,787

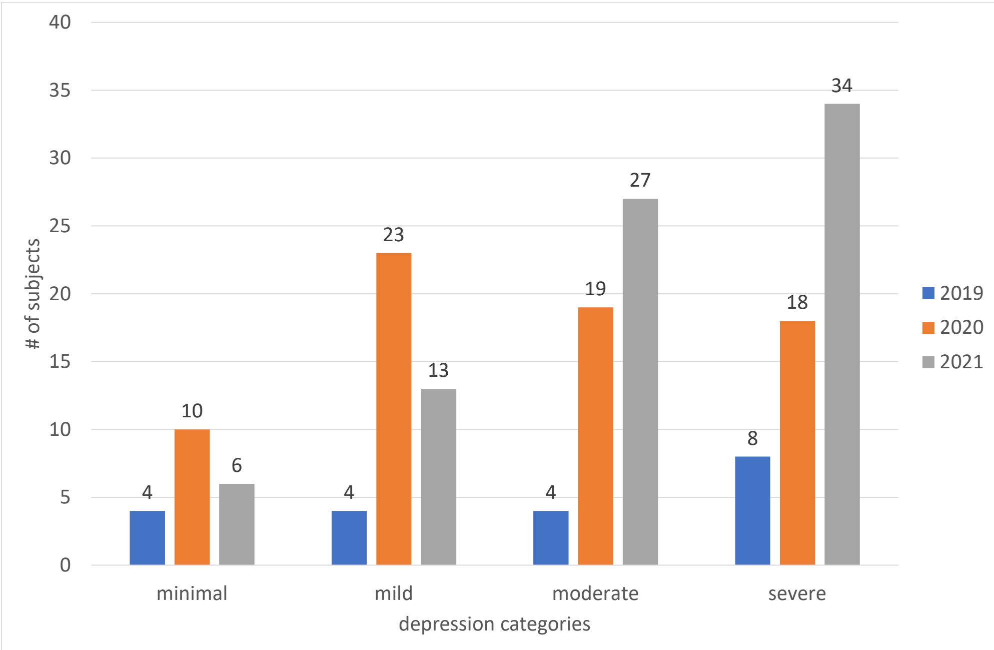

Figure 1: Statistics on eRisk subjects' depression categories.

4. Methods

Our approaches to leverage transfer learning are based on Bidirectional Encoder Representations from Transformers (BERT) [11]. The BERT model has been pre-trained on BookCorpus [39] and English Wikipedia on two objectives: Masked Language Model (MLM) [40] and Next Sentence Prediction [41, 42]. Furthermore, it is available in two different architectures:

- $\text{BERT}_{\text{BASE}}$ (number of layers=12, hidden size=768, number of self-attention heads=12)
- $\text{BERT}_{\text{LARGE}}$ (number of layers=24, hidden size=1024, number of self-attention heads=16)

Table 2
Experiments with SBERT models using the eRisk 2019-20 datasets. The names of the models are encoded as follows: *(base model)-(architecture)-(data used for pre-training)-(optional: data used for fine-tuning)-(pooling strategy)*. The evaluation measures are defined in Section 5.

SBERT model	AHR	ACR	ADODL	DCHR
bert-base-nli-cls-token	24.13%	60.07%	74.92%	24.44%
bert-base-nli-max-tokens	**28.47%**	61.90%	80.00%	24.44%
bert-base-nli-mean-tokens	25.40%	59.70%	74.87%	17.78%
bert-base-nli-stsb-mean-tokens	27.67%	60.78%	76.72%	21.11%
bert-large-nli-cls-token	26.93%	61.66%	77.21%	21.11%
bert-large-nli-max-tokens	28.10%	59.19%	78.84%	22.22%
bert-large-nli-mean-tokens	26.98%	60.83%	77.44%	25.56%
distilbert-base-nli-stsb-mean-tokens	28.25%	61.57%	78.15%	25.56%
distilbert-base-nli-stsb-quora-ranking	26.88%	59.42%	79.91%	26.67%
roberta-base-nli-stsb-mean-tokens	28.10%	**64.78%**	79.52%	27.78%
roberta-large-nli-stsb-mean-tokens	26.98%	63.81%	**81.27%**	**36.67%**
best scores	28.47%	64.78%	81.27%	36.67%

Every input sequence to BERT consists of tokens derived from the WordPiece [43] algorithm. The key advantage of this language representation model is that it overcomes the unidirectionality constraint of the previous ones (e.g., OpenAI GPT [37] and GloVe [33]). Moreover, it has been proved effective for fine-tuning and feature-based approaches [11]. We examine both of these strategies for our proposed approaches for the task 3 of eRisk 2021.

The Robustly Optimized BERT Pretraining Approach (RoBERTa) [13] has been trained longer on extended sequences, with bigger batches, and over more data. More specifically, its pre-training corpus also includes CC-News (portion of the CommonCrawl News dataset [44]), OpenWebText [45], and Stories [46]). Furthermore, there are some modifications to the training procedure (dynamic masking instead of static, no NSP loss, large mini-batches, and larger byte-level Byte-Pair Encoding (BPE) [47]).

Sentence-BERT (SBERT) [12] is an adaptation of pre-trained BERT and RoBERTa networks aiming to capture better sentence embeddings. For this purpose, it adds a pooling operation to the output of these models. To draw conclusions about the most appropriate SBERT model, we evaluate the performance of various SBERT models with respect to the predictive model described in Section 4.1 using the eRisk 2019 and 2020 datasets. The results are presented in Table 2. Our findings led us to use in our approaches a SBERT pre-trained model based on $\text{RoBERTa}_{\text{LARGE}}$, which was pre-trained on the combination of the Stanford NLI [48] and Multi-Genre NLI [49] and then fine-tuned on the STS benchmark dataset [50], with a mean-pool layer on the output to map subjects' posts to a vector space.

Overall, we mainly propose three approaches:

1. Feature-based transfer learning without using any training data
2. Feature-based transfer learning in combination with machine learning classification
3. Transfer learning with fine-tuning

The details of these approaches are presented in the following subsections.

4.1. Feature-based transfer learning without using any training data

In this approach, we use the aforementioned SBERT pre-trained model (max input sequence=128 tokens, i.e. padding the shorter and truncating the end of the longer sequences) to get the vector representation of Reddit posts which belong to the eRisk 2021 subjects. Similarly, we encode the responses to the BDI questionnaire into embeddings.

Next, we map subjects to the same vector space by calculating the mean of each feature of the post embeddings. Finally, we compare the vector of each subject with the vectors of the possible responses to each question in order to select the one with the maximum cosine similarity. The flowchart of this approach is shown in Figure 2.

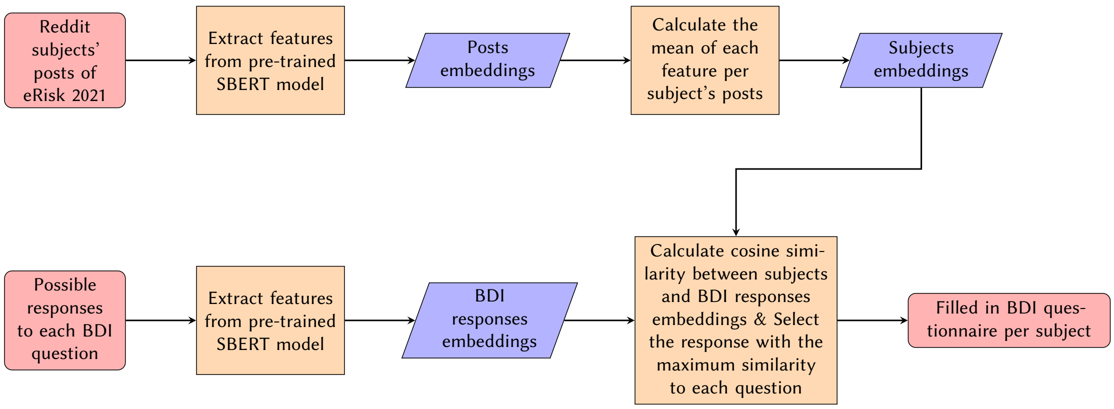

Figure 2: A flowchart of feature-based transfer learning approach without using any training data.

4.2. Feature-based transfer learning in combination with machine learning classification

This approach is quite similar to the previous one. We initially follow the same procedure to get the eRisk 2019 and 2020 subjects embeddings. However, this time, we use them as a training set to perform machine learning classification. The target variables for each subject are the 21 values (varying from 0–3 or 0–6 depending on the question) corresponding with his/her responses to the BDI questionnaire. Then, we apply the eRisk 2021 subjects embeddings as input to the derived trained model to make our predictions by filling in the BDI questionnaire per subject. The flowchart of this approach is shown in Figure 3.

The best-performing machine learning algorithms for this approach were selected utilizing the eRisk 2019 and 2020 datasets and using 10-fold cross-validation. Our experiments with various known classifiers are shown in Table 3. Slightly superior results are achieved with the AdaBoost [51], Linear SVM [52], and Naive Bayes [53] classifiers. Thus, we decided to employ the former two for our submitted runs.

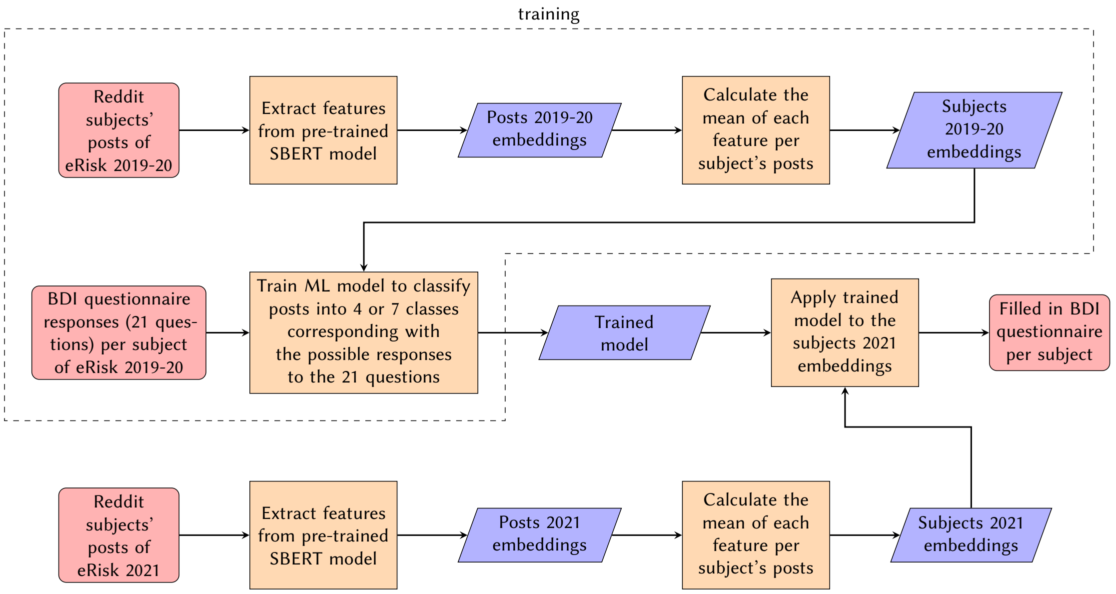

Figure 3: A flowchart of feature-based transfer learning approach in combination with machine learning classification.

Table 3
Experiments with various known classifiers using eRisk 2019-20 datasets and 10-fold cross-validation. Evaluation metrics are defined in Section 5.

Classifier	AHR	ACR	ADODL	DCHR
Nearest Neighbors	35.29%	68.14%	78.64%	28.33%
Linear SVM	**40.95%**	**72.00%**	80.64%	28.00%
RBF SVM	38.75%	69.81%	76.87%	20.76%
Gaussian Process	34.16%	67.76%	80.82%	26.67%
Decision Tree	31.79%	66.10%	81.91%	25.67%
Random Forest	39.19%	71.24%	81.55%	30.67%
Neural Net	36.59%	70.10%	82.48%	34.67%
AdaBoost	35.17%	68.76%	**83.21%**	35.00%
Naive Bayes	37.20%	69.67%	82.34%	**36.67%**
best scores	40.95%	72.00%	83.21%	36.67%

4.3. Transfer learning with fine-tuning

In this approach, we employ the eRisk 2019 and 2020 datasets as training set to fine-tune (epochs=3, batch size=32, and learning rate=2e-5) the RoBERTa$_{\text{BASE}}$ pre-trained model (max input sequence=128 tokens, i.e. padding the shorter and truncating the end of the longer sequences) to a classification task. To this end, we assigned subjects' BDI responses to each of their posts as target variables. A different fine-tuned model derived for each question, that outputs for each post of eRisk 2021 the probability of being related to each response.

Finally, in order to make the transition from post-level to subject-level, we apply two different

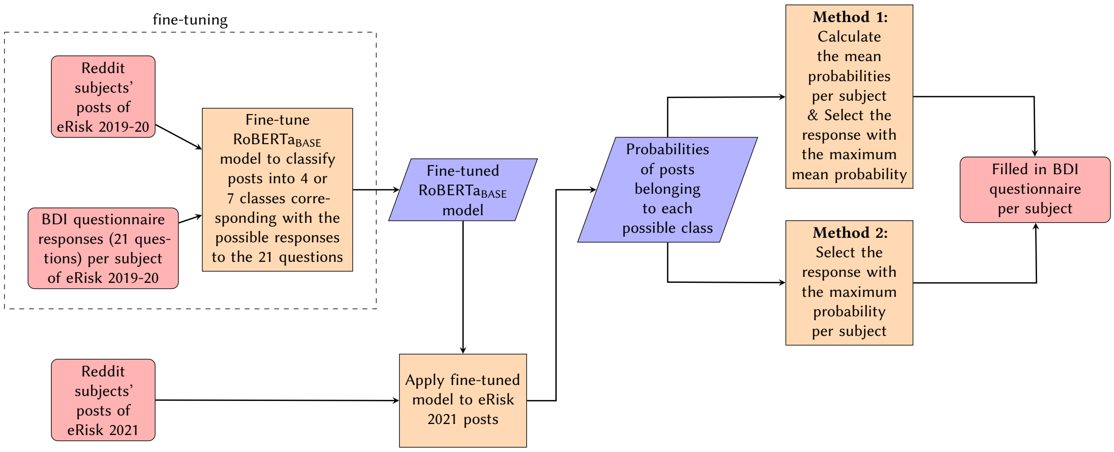

Figure 4: A flowchart of fine-tuning approach.

methods. In the first method, we calculate the mean probabilities per subject and select the response with the maximum mean probability, while in the second one, we simply select the response with the maximum probability per subject. The flowchart of this approach is shown in Figure 4.

5. Evaluation

In order to determine the subjects' depression level based on the BDI questionnaire, the responses to each question (out of 21 questions in total) are associated with integer values (i.e., 0–3). The sum of these 21 values is used to determine the depression level of a subject. The depression categories are associated with the depression levels in the following way:

- Minimal depression (depression levels 0–9)
- Mild depression (depression levels 10–18)
- Moderate depression (depression levels 19–29)
- Severe depression (depression levels 30–63)

The evaluation measures used by the organizers of this eRisk task to assess the performance of the submitted runs are as follows:

- Average Hit Rate (AHR): Reflects the accuracy of the responses to the BDI questionnaire submitted by the participants.
- Average Closeness Rate (ACR): Captures the deviation of the submitted responses from the real ones.
- Average Difference between Overall Depression Levels (ADODL): Captures the deviation of the sum of response values from the actual sum.
- Depression Category Hit Rate (DCHR): Reflects the accuracy of the depression category resulting from the sum of the submitted responses.

Table 4
Evaluation of DUTH-ATHENA's submissions. The best result across all participants for each measure is shown in the last line for comparison.

Run	Approach	AHR	ACR	ADODL	DCHR
DUTH_ATHENA MaxFT	3rd	31.43%	64.86%	74.46%	15.00%
DUTH_ATHENA MeanFT	3rd	32.02%	65.63%	73.81%	12.50%
DUTH_ATHENA MeanPosts	1st	25.06%	63.97%	80.28%	**30.00%**
DUTH_ATHENA MeanPostsAB	2nd	33.04%	**67.86%**	**80.32%**	27.50%
DUTH_ATHENA MeanPostsSVM	2nd	**35.36%**	67.18%	73.97%	15.00%
best scores		**35.36%**	73.17%	83.59%	41.25%

We also used the aforementioned measures to evaluate our experiments in the eRisk 2019 and 2020 datasets. Thus, we came up with the proposed runs for the first two approaches (Section 4.1 and 4.2). The third approach was quite time-consuming due to high computational cost and we could not afford to evaluate our methods.

5.1. Results

The evaluation of DUTH-ATHENA's submissions on the eRisk 2021 Task 3 are shown in Table 4. The second approach with the SVM classifier (`MeanPostsSVM`) achieved the highest score in terms of AHR, which is the most stringent measure. The same approach with the AdaBoost classifier (`MeanPostsAB`) ranked third among the 35 runs in ACR and ADODL.

Another promising finding was that the first approach (`MeanPosts`) performed well on predicting the depression levels and categories. In fact, this run ranked fourth in ADODL and DCHR among all submissions and first in DCHR among ours. This is remarkable since no annotated, task-dedicated data was used and the computational cost and execution time were the lowest among our runs. Finally, regarding the third approach with fine-tuning (`MaxFT` and `MeanFT`), the results are not comparable with the other two approaches because we utilized a smaller model architecture, due to computational limitations from our side, even thought we were expecting poorer results [13].

6. Conclusion

This paper presented our transfer learning approaches submitted to eRisk 2021 Task 3 utilizing a BERT-based pre-trained language model for a classification task that aims to automatically fill a depression questionnaire. The approaches utilize both feature-based and fine-tuning strategies. While our proposed models did not achieve high scores on all evaluation measures, we observed that this is a widespread problem among most of the submissions of the participants that maybe reflects the difficulty of the task. The modest performance of our third approach may be a result of the smaller model architecture or the matching of the subject's ground truth with all of their posts.

Nevertheless, we found that feature extraction from BERT-based pre-trained models achieved the best accuracy compared to the other participants' approaches. This suggests that further research in this direction could lead to promising outcomes. Future research should consider this potential more carefully, for example, experimenting with more and state-of-the-art pre-trained language models, such as Big Bird [54], and/or even more machine learning classifiers.

References

[1] A. Perrin, Social Media Usage: 2005-2015, Pew Research Center, 2015. URL: http://www.pewinternet.org/2015/10/08/2015/Social-Networking-Usage-2005-2015/.

[2] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, H. A. Schwartz, Facebook language predicts depression in medical records, Proceedings of the National Academy of Sciences 115 (2018) 11203–11208. doi:10.1073/pnas.1802331115, iSBN: 9781802331110 Publisher: National Academy of Sciences Section: Social Sciences.

[3] A. H. Orabi, P. Buddhitha, M. H. Orabi, D. Inkpen, Deep learning for depression detection of twitter users, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 88–97.

[4] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2968–2978. doi:10.18653/v1/D17-1322.

[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2021: Pathological Gambling, Self-harm and Depression Challenges, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 650–656. doi:10.1007/978-3-030-72240-1_76.

[6] 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, Springer International Publishing, Cham, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[7] D. E. Losada, F. Crestani, J. Parapar, eRisk 2020: Self-harm and Depression Challenges, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 557–563. doi:10.1007/978-3-030-45442-5_72.

[8] 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. doi:10.1001/archpsyc.1961.01710120031004.

[9] 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.), Conference and Labs of the Evaluation Forum (CLEF), 2019, p. 21.

[10] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2020: Early risk prediction on the internet, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma,

C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, 11th International Conference of the CLEF Association, CLEF 2020, Springer International Publishing, Cham, 2020, pp. 272–287.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
[12] 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410.
[13] 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, 2019. arXiv:1907.11692.
[14] A. Trifan, P. Salgado, L. Oliveira, BioInfo@UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Neveol (Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696, Thessaloniki, Greece, 2020, p. 11.
[15] S. G. Burdisso, M. Errecalde, M. Montes-y Gomez, 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, volume 2380, Lugano, Switzerland, 2019, p. 18.
[16] D. Maupome, M. D. Armstrong, R. Belbahar, J. Alezot, R. Balassiano, M. Queudot, S. Mosser, M.-J. Meurs, Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Neveol (Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696, Thessaloniki, Greece, 2020, p. 13.
[17] A. Madani, F. Boumahdi, A. Boukenaoui, C. Kritli, H. Hentabli, USDB at eRisk 2020: Deep learning models to measure the Severity of the Signs of Depression using Reddit Posts, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Neveol (Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696, Thessaloniki, Greece, 2020, p. 9.
[18] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197. doi:10.1016/j.eswa.2019.05.023.
[19] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.
[20] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, M. Zhu, A Practical Algorithm for Topic Modeling with Provable Guarantees, in: S. Dasgupta, D. McAllester (Eds.), Proceedings of the 30th International Conference on Machine Learning, volume 28(2), PMLR, 2013, pp. 280–288.

[21] L. Xian, S. D. Vickers, A. L. Giordano, J. Lee, I. K. Kim, L. Ramaswamy, #selfharm on Instagram: Quantitative Analysis and Classification of Non-Suicidal Self-Injury, in: 2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI), 2019, pp. 61–70. doi:10.1109/CogMI48466.2019.00017.

[22] 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. doi:10.3115/v1/P15-1162.

[23] D. Maupomé, M. Queudot, M.-J. Meurs, Inter and intra document attention for depression risk assessment, in: Canadian Conference on Artificial Intelligence, Springer, 2019, pp. 333–341.

[24] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in: The handbook of brain theory and neural networks, MIT Press, Cambridge, MA, USA, 1998, pp. 255–258.

[25] J. L. Elman, Finding structure in time, Cognitive Science 14 (1990) 179–211. doi:10.1016/0364-0213(90)90002-E.

[26] A. Graves, J. Schmidhuber, Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, Advances in Neural Information Processing Systems 21 (2008).

[27] C. Baziotis, N. Pelekis, C. Doulkeridis, DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 747–754. doi:10.18653/v1/S17-2126.

[28] Y. Zhang, J. Wang, X. Zhang, YNU-HPCC at SemEval-2018 Task 1: BiLSTM with Attention based Sentiment Analysis for Affect in Tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 273–278. doi:10.18653/v1/S18-1040.

[29] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 207–212. doi:10.18653/v1/P16-2034.

[30] J. Turian, L.-A. Ratinov, Y. Bengio, Word Representations: A Simple and General Method for Semi-Supervised Learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 384–394.

[31] 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: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, volume 2125, Avignon, France, 2018, p. 15.

[32] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Neveol

(Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696, Thessaloniki, Greece, 2020, p. 12.
[33] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.
[34] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. doi:10.18653/v1/D18-2029.
[35] P. Abed-Esfahani, D. Howard, M. Maslej, S. Patel, V. Mann, S. Goegan, L. French, Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, volume 2380, Lugano, Switzerland, 2019, p. 9.
[36] R. Martınez-Castano, A. Htait, L. Azzopardi, Y. Moshfeghi, Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Neveol (Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, volume 2696, Thessaloniki, Greece, 2020, p. 16.
[37] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-training, 2018. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
[38] A. Conneau, G. Lample, Cross-lingual language model pretraining, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems (NeurIPS 2019), volume 32, Curran Associates, Inc., Vancouver, Canada, 2019, p. 11.
[39] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, IEEE Computer Society, USA, 2015, p. 19–27. doi:10.1109/ICCV.2015.11.
[40] W. L. Taylor, "Cloze Procedure": A New Tool for Measuring Readability, Journalism Quarterly 30 (1953) 415–433. doi:10.1177/107769905303000401.
[41] Y. Jernite, S. R. Bowman, D. Sontag, Discourse-based objectives for fast unsupervised sentence representation learning, 2017. arXiv:1705.00557v1.
[42] L. Logeswaran, H. Lee, An efficient framework for learning sentence representations, in: Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 2018, p. 16.
[43] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, CoRR

abs/1609.08144 (2016).
[44] S. Nagel, CC-News, https://commoncrawl.org/2016/10/news-dataset-available/, 2016.
[45] A. Gokaslan, V. Cohen, Openwebtext corpus, http://Skylion007.github.io/OpenWebTextCorpus, 2019.
[46] T. H. Trinh, Q. V. Le, A simple method for commonsense reasoning, 2019. arXiv:1806.02847.
[47] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.
[48] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Lisbon, Portugal, 2015, p. 632–642.
[49] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. doi:10.18653/v1/N18-1101.
[50] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1–14. doi:10.18653/v1/S17-2001.
[51] Y. Freund, R. E. Schapire, A short introduction to boosting, in: In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1999, pp. 1401–1406.
[52] N. Cristianini, E. Ricci, Support Vector Machines, in: M.-Y. Kao (Ed.), Encyclopedia of Algorithms, Springer US, Boston, MA, 2008, pp. 928–932. doi:10.1007/978-0-387-30162-4_415.
[53] G. I. Webb, Naïve Bayes, in: C. Sammut, G. I. Webb (Eds.), Encyclopedia of Machine Learning, Springer US, Boston, MA, 2010, pp. 713–714. doi:10.1007/978-0-387-30164-8_576.
[54] M. Zaheer, G. P. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. M. Pham, A. Ravula, Q. Wang, L. Yang, A. M. E. H. Ahmed, Big Bird: Transformers for Longer Sequences, in: Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, 2020, p. 15.