CeDRI at eRisk 2021: A Naive Approach to Early Detection of Psychological Disorders in Social Media

Rui Pedro Lopes[1]

[1]Research Center for Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Portugal

Abstract

This paper describes the participation of the CeDRI team in eRisk 2021 tasks, particularly, the Task 1: Early Detection of Signs of Pathological Gambling and Task 2: Early Detection of Signs of Self-Harm. The main difference between these two is that the first is a "test only" challenge, where no training data is supplied. The second task has labeled data available, which can be used for training. Both tasks were addressed using the same algorithms, using a custom training set for Task 1 and the provided data in the second. The algorithms were TfIdf vectorizer with a Logistic Regression layer, Word2Vec vectorizer with LSTM and Word2Vec vectorizer with CNN. All vectorizers and Neural Networks were trained solely with the training data. As expected, the algorithms did not state-of-the-art, but the experience allowed to reflect in several aspects related to the importance of proper dataset preparation and processing.

Keywords

Early Risk Detection, Tf-Idf, Word2Vec, Recursive Neural Networks, Dataset Heuristics, DL4J.

1. Introduction

The term social network refers to a person's connections to other people. In fact, creating and maintaining social networks provide opportunities to connect with others who have similar interests. Although initially applied in the context of "real-world" or physical, the concept expanded to also include platforms that support online communication, such as Instagram, Twitter or Reddit. Digital platforms further enhance these opportunities, allowing forming relationships with people never met in person. Geographical barriers are attenuated or eliminated, allowing to actively engage with people around the world. They can explore their curiosity, pick up hobbies, or just spend time online. The possibility to write, participate or communicate without restrictions also provides a means to unburden or receive emotional support. Some people resort to social networks to talk about their state of mind, their feelings, distresses and other problems.

In opposition to verbal and direct communication, the content available in the social networks is persistent, allowing asynchronous access data and providing a good means for psychological and health related studies and analysis [1, 2, 3]. According to several findings, people's mental state can be inferred from their social networks narratives [4, 5]. Based in this, the CLEF eRisk challenges harness this opportunity to explore issues of evaluation methodologies, performance

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
rlopes@ipb.pt (R. P. Lopes)
0000-0002-9170-5078 (R. P. Lopes)
© 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)

CEUR-WS.org/Vol-2936/paper-80.pdf

metrics and other aspects related to building test collections and defining challenges for early risk detection [6, 7, 8, 9].

This year's challenge has three tasks. Task 1, on early risk detection of pathological gambling, and Task 2, on early risk detection of self-harm, consist of sequentially processing pieces of evidence and detect early traces of pathological gambling and self-harm , respectively, as soon as possible. Task 3, measuring the severity of the signs of depression, consists of estimating the level of depression from a thread of user submissions. The CeDRI team participated in Task 1 and Task 2, where users' posts are processed in the same order in which they are sent, to chronologically monitor the users' activity.

This paper presents the participation of the CeDRI team in the pathological gambling and in the self-harm early detection challenges of CLEF 2021. In task 1, two runs where executed, using a Long-short Term Memory (LSTM) and Convolutional Neural Network (CNN) deep neural networks, both with Word2Vec embeddings. Task 2 used three runs, with LSTM, CNN with Word2Vec embeddings, like the previous task, and a logistic regression layer with Tf-Idf vectorizer. Although the results were very close within the runs, the best results in Task 1 was *latency-weighted F1=0.141* (with the LSTM) and in Task 2 *latency-weighted F1=0.206* (with the CNN).

The rest of the paper is organized as follows. Section 2 covers the considerations regarding the datasets, while section 3 introduces the proposed method. Analysis of the results of experiments are presented in section 4 and finally, the conclusion and suggested directions for future works are presented in section 5.

2. Dataset

The machine learning area is characterized by three main approaches of learning [10]:

- supervised - maps an input to an output based on example input-output pairs;
- unsupervised - patterns are learned without any explicit feedback;
- reinforcement - learns from a series of reinforcements, such as rewards and punishments.

These are applied in several areas and with several purposes, such as classification, prediction, estimation, affinity grouping, clustering and profiling. The eRisk challenge Task 1 and 2 is mainly a classification problem, widely approached with supervised learning methods. In these problems, a learning agent is shown what to do through an annotated set of training examples, and it is expect an automated learning algorithm to generalize from these examples.

For this, it is fundamental to understand and make sure that the training data is adequate and it is well labeled.

2.1. Text pre-processing

Social networks' posts often include tokens that do not represent words, such as URLs, HTML entities, users' handles, or others. Some of these do not bring relevant information to infer the psychological condition of the user and may affect the performance of classification. The pre-processing applied in both tasks included the following operations:

- unescape html entities (ex: < or <)
- remove handles (@abcd @pqrs)
- remove URLs (https://erisk.irlab.org)
- normalize lengthening (111111 -> 11; kkkkkkkkkkk -> kk)
- remove numbers
- convert to lowercase (Tomorrow -> tomorrow)
- strip punctuation
- tokenize
- perform stemming

The vocabulary is substantially reduced, as well as the word variations (Table 1). The same pre-processing approach was applied in both tasks (sections 2.2 and 2.3).

Table 1
Pre-processing sample.

Original text		Pre-processed text
We will be having our next meeting this evening at 5:00pm EST (9:00pm GMT). Meetings are 1 hour. Participants must use Skype audio and video. If you'd like to join, [DM me](http://www.reddit.com/message/compose/?to=JeffW55&subject=ProblemGamblingSupportGroup) with your Skype name so you can be added to the call. Thanks. Jeff	→	[next, meet, even, pm, pm, gmt, meet, hour, particip, must, skype, audio, video, you'd, like, join, me, gambl, support, group, skype, name, ad, call, thank, jeff]

2.2. Task 1: pathological gambling dataset

The challenge consists of sequentially processing pieces of evidence and detect early traces of pathological gambling signs in texts written in Social Media. This was an "only test" task, so no training data was provided. The test collection format is a collection of writings (posts or comments) from a set of Social Media users, labeling two categories of users, pathological gamblers and non-pathological gamblers, and, for each user, the collection contains a sequence of writings (in chronological order) [11].

Since the challenge did not provide labeled data, a custom dataset, based on Reddit, was built. For that, the Python Pushshift.io API Wrapper (PSAW - https://github.com/dmarx/psaw) was used to retrieve posts from the Pushshift initiative (https://pushshift.io), in Comma Separated Values (CSV) format. This allowed to remove the limit of 1000 posts that could be downloaded from Reddit directly. The dataset was built based on the `r/GamblingAddiction` and `r/problemgambling` communities. In addition, a random set of posts was also downloaded to complement the dataset with non-gambling related content (Table 2).

There is a considerable number of posts available after downloading, in a total of 73064 referring gambling issues and 47103 posts of random subjects. However, extracting data from the CSV files failed in many posts, having only 7079 posts and 2306, respectively. This was due to incompatibility issues between the post text and the CSV encoding, related to the appearance of commas (',') in the text and unterminated '"', which made the issue of extracting the columns

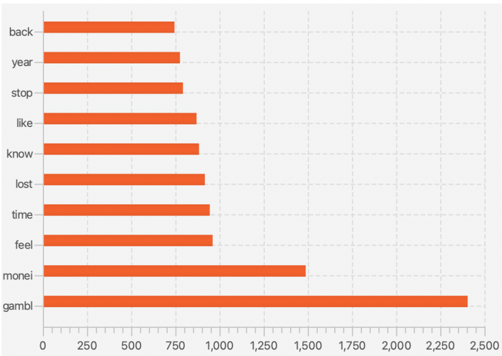
Table 2
Summary of the training data set for eRisk 2021 Pathological gambling task

Reddit Community	Number of Posts	Usable Posts	Dataset	Label
`r/GamblingAddiction`	16528	1467	1467	True
`r/problemgambling`	56536	5612	839	True
random	47103	2306	2306	False

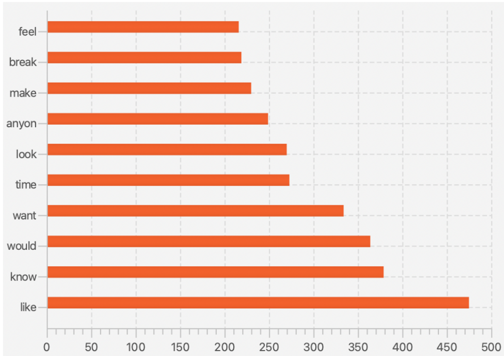
very difficult and error sensitive. Because of balancing issues, the dataset was build with 2306 posts labeled with `False` and 2306 posts with `True`.

Each post was stored in a single file, prefixed with `pos` or `neg` followed by a number (e.g. `pos_1762.txt`, `neg_2032.txt`). It was decided not to associate or track the users, so each post is individual and not related to any other.

After building the dataset, the most frequent tokens in the gambling related posts (1a) and in the non-gambling related posts (1b) were calculated (Figure 1). As expected, tokens like `gambi`, `monei`, or `stop` appear in the vocabulary for gambling posts. For random, `like`, `know` and `would` are very frequent.



(a) gambling related posts.



(b) Non-gambling related posts.

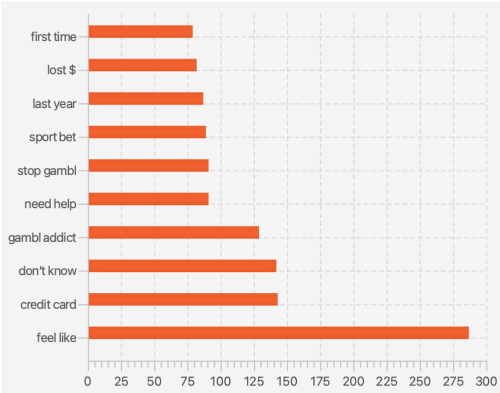
Figure 1: The ten most frequent words.

Next, the same operation was performed for bi-grams, to better understand the context of the words (Figure 2).

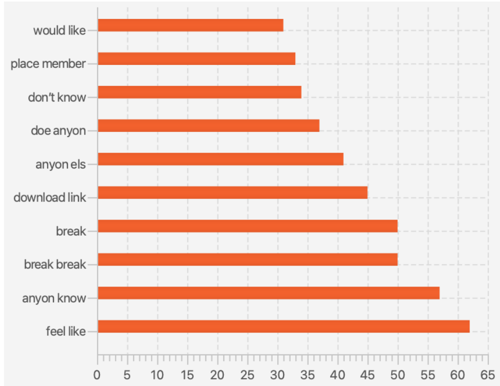
In these, `feel like` is transversal to both types of posts, although `credit card` and `gambli addict`, for example, are clearly indicating the type of posts.

2.3. Task 2: self-harm dataset

The training data provided XML files for 340 subjects, 41 of which belonging to the self-harm group, 299 to the control group (Table 3). The total number of writings in the self-harm group is 7,192 posts in contrast to 163,506 in the control group. The difference between the groups is also very significant in the average number of writings per subject: 175.4 in the self-harm group and 546.8 in the control group. The average length of the users' (subjects') writings is



(a) Gambling related posts.



(b) Non-gambling related posts.

Figure 2: The ten most frequent bi-grams.

179.1 and 129.2 respectively, and the number of tokens is 15.12 and 10.6. Although the control subjects write more posts, they are, in average, shorter. The dataset is also provided with the test writing, in the same format. They are also present in table 3, for completeness.

Table 3
Summary of the data set for eRisk 2021 Self-harm task

	Train		Test		Full	
	Self-harm	Control	Self-harm	Control	Self-harm	Control
Subjects	41	299	104	319	145	618
Min Posts	8	10	9	9	8	0
Max Posts	997	1992	942	1990	997	1992
Total Posts	7192	163506	11691	92146	18883	255652
Avg Posts	175,4	546,8	112,4	288,9	130,2	413,68
Min Length	0	0	0	0	0	0
Max Length	5880	54796	6627	56651	6627	56651
Total Length	1288542	21129774	1823906	7339145	3112448	28468919
Avg Length	179,1	129,2	156	79,6	164,8	111,4
Min Tokens	0	0	0	0	0	0
Max Tokens	546	3342	559	1334	559	3342
Total Tokens	108752	1730021	148204	568180	256956	2298201
Avg Tokens	15,12	10,6	12,7	6,2	13,6	9

In addition, not all posts are of the same language. Using OpenNLP's language detection model, a total of 81 different languages were counted. Table 4 show the 15 more frequent languages within the writings.

The dataset uses binary labels on the subjects, as having (positive) and not-having (negative) self-harm (ground truth). As seen in table 3, each subject has an arbitrary number of posts, and it is not expected that all of them will be strictly related to whether an user self-harms or not. The main approach in this work, is to use a machine learning approach that uses text to

Table 4
Different languages in the training set

Code	Language	Count
eng	English	107123
tur	Turkish	44288
cmn	Chamic languages	2353
war	Waray	1625
lat	Latin	1423
min	Minangkabau	1385
plt	Pali	1123
afr	Afrikaans	1081
vol	Volapük	983
mri	Mossi	973
por	Portuguese	781
epo	Esperanto	596
nob	Norwegian Bokmål	499
ron	Romany	391
ceb	Cebuano	364

predict whether a message belongs to a positive or negative user, so the classifier should not be trained with just the ground truth. Some selection on the posts have to be made, so that only the self-harm related writings are kept as positive samples in the training set.

Based on Non-Suicidal Self-Injury (NSSI) words [12], a selection was made on the posts to extract individual writings to be used as positive examples. The examples were written in two directories (`pos/` for positive and `neg/` for negative) with the following name schema: `subject280_2.txt`, where the first number is the subject number and the second is this subject's post number. After selecting writings based on NSSI words, and excluding all languages except English, a total of 391 positive labeled writings remained. For balance, the same number of negative labeled writings were selected.

The tokens frequency were also extracted from both the positive and the negative writings. In this case, the bi-grams (Figure 3) and tri-grams (Figure 4) are presented.

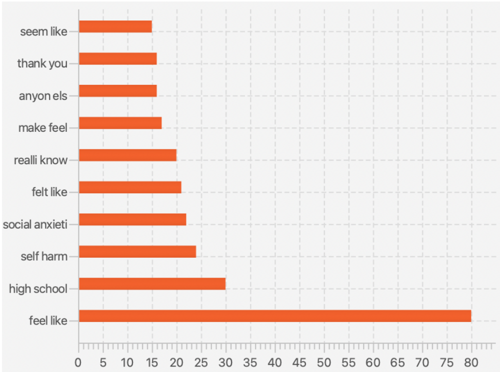
3. Proposed methods

This section presents the models and experiments conducted for the eRisk 2021 task. First, the classification methods require that text be converted to vectors.

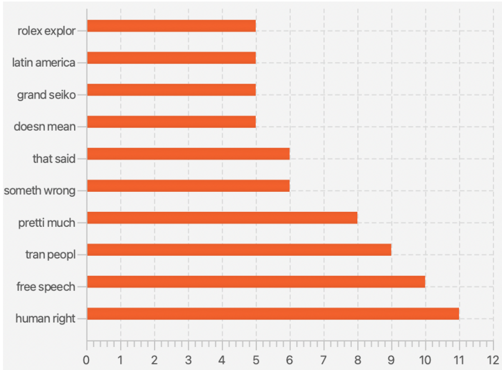
3.1. Vectorizers

All the methods rely on the vectorization of the subjects' writings. Two vectorizers were trained, based on TfIdf and Word2Vec, both with the same text pre-processing techniques (section 2.1).

- TfIdf:
  - minimum word frequency = 2;

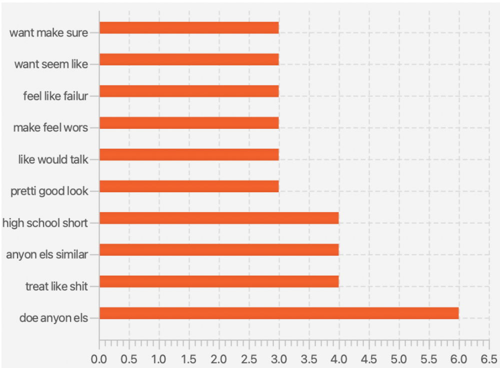


(a) Self-harm related posts.

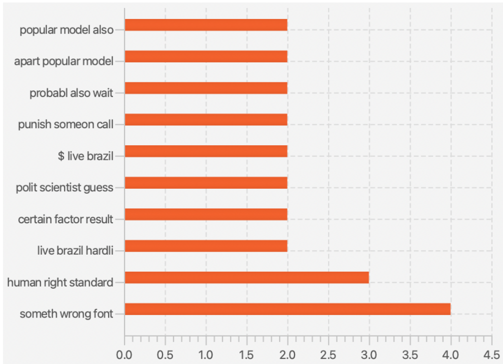


(b) Non-self-harm related posts.

Figure 3: The ten most frequent bi-grams.



(a) Self-harm related posts.



(b) Non-self-harm related posts.

Figure 4: The ten most frequent tri-grams.

- Word2Vec:
  - minimum word frequency = 5;
  - number of iterations = 1;
  - number of epochs = 5;
  - layer size = 128;
  - window size = 5;

3.2. Classifiers

Three classification models were build for the tasks. The first is a simple Logistic Regression layer using the TfIdf vectorizer, used in both Tasks 1 and 2:

- output dimension = 2;
- weight initialization algorithm = XAVIER;

- activation = SOFTMAX;
- optimization algorithm = STOCHASTIC_GRADIENT_DESCENT;
- updater = Nesterovs(0.1, 0.9)
- batch size = 32;

Another classifier was built using a CNN with Word2Vec vectors as input, used only in Task 2:

- weight initialization algorithm = RELU;
- activation = LEAKYRELU;
- updater = Adam(0.01);
- convolution mode = SAME;
- l2 = 0.0001;
- convolution layer 1 = [128, 100], kernel size = [3, 128]
- convolution layer 2 = [128, 100], kernel size = [4, 128]
- convolution layer 3 = [128, 100], kernel size = [5, 128]
- merge(cl1, cl2 and cl3)
- global pooling with dropout = 0.5
- loss function = MCXENT
- dense layer = [100, 2], activation = SOFTMAX

Finally, a classifier based on LSTM with Word2Vec vectors as input, used in both Tasks 1 and 2:

- updater = Adam(5e-3)
- l2 = 1e-5;
- weight initialization algorithm = XAVIER;
- lstm layer = [128, 256], activation = TANH);
- lstm output layer = [256, 2], activation = SOFTMAX; loss function = MCXENT

4. Analysis of the results

In task 1, according to the eRisk 2021 evaluation report, the maximum number of all users writings was 2000. Of these, only 271 were processed, in 1 day 5 hours, 44 minutes and 10 seconds, until the servers were shutdown. The unavailability, at the time, of an additional GPU made the processing time much slower and, as such, 1 day was not enough to process the whole set. Two runs were executed, based on LSTM and TfIdf (Table 5).

The final results are far from the best in all metrics. Nevertheless, the LSTM performed better, although marginally, than TfIdf, a much simpler classifier.

In task 2, the maximum number of all users writings were 1999. Of these, only 369 were processed, taking 1 day 9 hours, 51 minutes and 27 seconds. As before, and although a GPU was available in this task, the system was not able to process the totality of test users until the server was shutdown. Three runs were executed, based on LSTM, CNN and TfIdf (Table 6).

Table 5
Task 1 runs

Run	Method	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency - weightedF1$
0	LSTM	.076	1	.142	.079	.060	2	.996	.141
1	TfIdf	.070	1	.131	.066	.065	1	1	.131

Table 6
Task 2 runs

Run	Method	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency - weightedF1$
0	LSTM	.11	.993	.199	.109	.09	2	.996	.198
1	CNN	.116	1.0	.207	.113	.085	2	.996	.206
2	TfIdf	.105	1	.19	.096	.094	1	1.0	.19

It seemed that the CNN performed better in some metrics, although marginally, compared with LSTM, with TfIdf getting very low scores. Moreover, the algorithms seems to be highly inclined to emit positive decisions, with perfect recall but extremely low precision. Although it is not clear, this may be due to the fact that the posts are processed individually, without any consideration of the previous writings. Some window or accumulator approach could be used to understand if this is the issue.

Overall, the three methods can be improved. They were rather close, which gives the indication that the main issue is with the selection of the training dataset. A deeper understanding is necessary regarding the dataset and, after that, new methods can be devised and tested.

5. Conclusions

This paper describes the CeDRI submission to the CLEF eRisk 2021 task 1 and 2 on detecting early signs of pathological gambling and self-harm in social media posts. Three methods were presented that seek to classify each writing independently of the others using only information about the text. The first task is a "test only", so it was necessary to build a training set based on posts collected from Reddit. Task 2 required the processing and filtering of the writings in order to isolate the posts that refer to self-harm from the others, and use these for training the classifiers.

Due to the simple classifiers used, state-of-the-art results were not expected. The main purpose was to try to understand the effectiveness of building training sets based on simple heuristics filters. For future work, the inclusion of more features, such as Part of Speech (PoS) frequency, post date and time, and others should be studied.

Acknowledgments

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/05757/2020.

References

[1] D. Marengo, C. Montag, C. Sindermann, J. D. Elhai, M. Settanni, Examining the links between active Facebook use, received likes, self-esteem and happiness: A study using objective social media data, Telematics and Informatics 58 (2021) 101523. URL: https://linkinghub.elsevier.com/retrieve/pii/S0736585320301829. doi:10.1016/j.tele.2020.101523.

[2] L. Faelens, K. Hoorelbeke, B. Soenens, K. Van Gaeveren, L. De Marez, R. De Raedt, E. H. Koster, Social media use and well-being: A prospective experience-sampling study, Computers in Human Behavior 114 (2021) 106510. URL: https://linkinghub.elsevier.com/retrieve/pii/S0747563220302624. doi:10.1016/j.chb.2020.106510.

[3] X. Chen, Z. Pan, A review on assessment, early warning and auxiliary diagnosis of depression based on different modal data, in: Z. Pan, X. Hei (Eds.), Twelfth International Conference on Graphics and Image Processing (ICGIP 2020), SPIE, Xi'an, China, 2021, p. 75. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11720/2589413/A-review-on-assessment-early-warning-and-auxiliary-diagnosis-of/10.1117/12.2589413.full. doi:10.1117/12.2589413.

[4] B. Moulahi, J. Azé, S. Bringay, DARE to Care: A Context-Aware Framework to Track Suicidal Ideation on Social Media, in: A. Bouguettaya, Y. Gao, A. Klimenko, L. Chen, X. Zhang, F. Dzerzhinskiy, W. Jia, S. V. Klimenko, Q. Li (Eds.), Web Information Systems Engineering – WISE 2017, volume 10570, Springer International Publishing, Cham, 2017, pp. 346–353. URL: http://link.springer.com/10.1007/978-3-319-68786-5_28. doi:10.1007/978-3-319-68786-5_28, series Title: Lecture Notes in Computer Science.

[5] Z. Zhang, G. Bors, "Less is more": Mining useful features from Twitter user profiles for Twitter user classification in the public health domain, Online Information Review 44 (2019) 213–237. URL: https://www.emerald.com/insight/content/doi/10.1108/OIR-05-2019-0143/full/html. doi:10.1108/OIR-05-2019-0143.

[6] D. E. Losada, F. Crestani, J. Parapar, eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations, in: G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 10456, Springer International Publishing, Cham, 2017, pp. 346–360. URL: http://link.springer.com/10.1007/978-3-319-65813-1_30. doi:10.1007/978-3-319-65813-1_30, series Title: Lecture Notes in Computer Science.

[7] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview) (2018) 20.

[8] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019 Early Risk Prediction on the Internet, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 340–357. doi:10.1007/978-3-030-28577-7_27.

[9] 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, volume 12260, Springer International Publishing, Cham,

2020, pp. 272–287. URL: https://link.springer.com/10.1007/978-3-030-58219-7_20. doi:10.1007/978-3-030-58219-7_20, series Title: Lecture Notes in Computer Science.
[10] S. J. Russell, P. Norvig, Artificial intelligence: a modern approach, Pearson series in artificial intelligence, fourth edition ed., Pearson, Hoboken, 2021.
[11] D. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use, in: Proc. of Experimental IR Meets Multilinguality, Multimodality, and Interaction, 7th International Conference of the CLEF Association, CLEF 2016, Evora, Portugal, 2016, pp. 28–39.
[12] M. M. Greaves, C. Dykeman, A Corpus Linguistic Analysis of Public Reddit Blog Posts on Non-Suicidal Self-Injury, arXiv:1902.06689 [cs] (2019). URL: http://arxiv.org/abs/1902.06689, arXiv: 1902.06689.