A Baseline Approach for Early Detection of Signs of Anorexia and Self-harm in Reddit Posts

Nona Naderi[1,2,3], Julien Gobeil[1,2], Douglas Teodoro[1,2], Emilie Pasche[1,2], and Patrick Ruch[1,2]

[1] HES-SO/HEG Geneva, Information Sciences, Geneva, Switzerland
[2] SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland
[3] University of Toronto, Toronto, Canada
nona.naderi;julien.gobeill;douglas.teodoro;emilie.pasche;patrick.ruch@hesge.ch

Abstract. This paper describes the systems developed by the BiTeM team for the CLEF eRisk Task 1 and 2, 2019. The goal was to predict the risk of anorexia and self-harm from user-generated content on Reddit. Several approaches based on supervised learning were used to estimate the risk of anorexia and self-harm. The systems were able to achieve low to moderate results.

Keywords: Natural Language Processing · Text Mining · Mental health.

1 Introduction

This paper describes the participation of BiTeM group at CLEF 2019 eRisk early risk detection of anorexia and signs of self-harm (T1 and T2, respectively) on users of the Reddit community. Reddit[4] is a community-driven platform that consists of various "subreddits" on different topics, and users post contents, such as images and texts, or comment on other posts. The objective of tasks T1 and T2 of CLEF eRisk 2019 was to predict early signs of anorexia and self-harm, respectively, among Reddit users. Given a sequence of posts from users published over a period of time, the system should be able to detect as early as possible whether a user is showing signs of anorexia (T1) or self-harm (T2). We first describe our submission models that are based on bag-of-words, and then explore the additional models based on mutual information and convolutional neural networks, and an ensemble model that used all three methods. The results presented here include the official and post-competition runs. We further describe our findings on these tasks and suggest possible future improvements.

[4] https://www.reddit.com/

2 Related work

Prediction of early traces of risks from individuals' generated content has received substantial attention in recent years due to the introduction of competitions, such as CLEF (2017-2019) [6–8], CLPsych Shared Task (2015-2019) [2, 11, 9, 21] and the Audio/Visual Emotion Recognition (AVEC) Depression Subchallenge (2013-2017) [20, 19, 18, 15]. These shared tasks focus on identifying various mental disorders from different types of content, such as depression from Reddit posts [6–8], depression and PTSD from tweets [2], the degree of distress from Reachout forum posts [11], suicide risk using Reddit posts [21], depression scale from audio, visual and text of interview responses [15], and anorexia and self-harm from Reddit posts [8]. Previous studies have shown that by looking at one's written texts on social media, we may learn more about the mental and social state of that individual [1, 3, 16]. More generally, letters or diaries have been used as material to study different aspects of human behaviours by researchers in social and health sciences [12]. In response to the prediction of anorexia in CLEF 2018, promising results were achieved by an ensemble model using two CNN models and a logistic regression model trained with Bag of Words and metadata features [17].

3 Datasets

For the early risk detection of anorexia (T1), the training dataset was based on the eRisk 2018 data [7], which contained a history of writings from social media users. The dataset specified whether an individual was diagnosed with anorexia or not, but it did not say which writings of that user indicate signs of anorexia. Table 1 shows the statistics of the training set. For the early detection of signs of self-harm (T2), no training set was provided.

Table 1. Characteristics of training set of anorexia.

	Anorexia
Users	471
Posts+comments	207,604
Avg documents per user	441
Risk alert/non alert	410/61

The source of test set was also based on the data provided in eRisk 2017 and 2018, but it was released item by item, through a news feed simulation server provided by the task organisers, and the participant systems were supposed to determine the risk signs of anorexia or self-harm as early as possible. A classifier would then retrieve a post stream, containing several users posts, classify them and submit the results to the server. Then, it could retrieve the next stream of posts. Each task provided around 2000 post streams. We used 2000 chunks for

T1 and 1992 chunks for T2. Table 2 shows that the statistics of the retrieved posts for both tasks.

Table 2. Characteristics of test sets.

	Anorexia	self-harm
Users	815	340
Comments	391,551	120,935
Posts	178,915	49,753
Avg documents per user	700	502
Risk alert/non alert	73/742	41/299

4 Methods

We explored several approaches to predict the early signs of anorexia and self-harm among Reddit users. Due to the lack of training data at the post level, we investigated data-driven approaches, leveraging on the large set of Reddit data. Our hypothesis was that we could use the high-level classes, provided by the subreddits, to derive supervised or semi-supervised collections for training our models.

4.1 Model 1 - Bag-of-words model

As a first round of experiments, we exploited the 2018 competition anorexia task dataset. In 2018, the data consisted of ten chunks of several posts aggregated for each user. One key factor is that relevant judgments (positive or negative) were provided for each user, not per chunk. We thus decided to aggregate all posts of a given user into a unique virtual document. For task 1, run 0, we trained a support vector machine (SVM) model with linear kernel[5][13] using the 2018 training data and *tf-idf* representation, and evaluated it on 2018 test data. The test chunks for each user, were aggregated at each round. In this configuration, our SVM model reached a promising F-measure of 0.72. For the other runs of task 1, we used Reddit posts to train our models as described below.

For task 2, on the other hand, no training data was provided. Thus, we relied on Reddit posts to identify the vocabulary employed by users that write about their problems using three years of Reddit posts ($\sim$ 1TB of data) that we downloaded. We used posts from the r/selfharm subreddit as positive training data. For negative data, our impressions on the 2018 data was that negative posts dealt with general subjects, such as links to videos or gaming. This impression was reinforced by the visualization of the most informative features for the SVM model (see Figure1 below). For example, calories, anorexia, weight, help, and fat were among the most informative features for the positive class and https,

[5] We used scikit-learn, https://scikit-learn.org

lol, game, show, and www were among the informative features for the negative class.

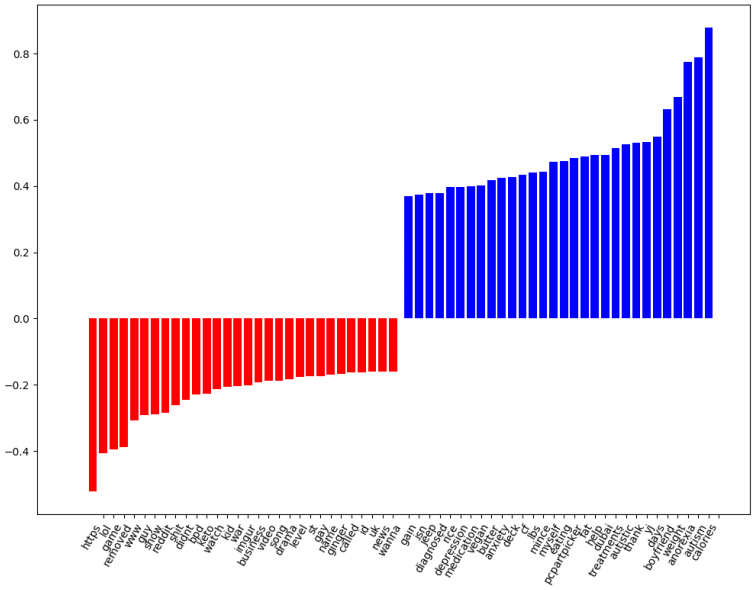

Fig. 1. The thirty most important features for SVM choice on 2018 eRisk anorexia dataset. Features related to positive instances are in blue, features related to negative instances are in red.

We thus randomly sampled the posts in non-positive subreddits for extracting negative data. Training datasets were then designed, and we applied SVM with the same settings on them. We used a similar approach for task 1, leveraging the r/EatingDisorder subreddit as positive training data. The last setting was to decide how many Reddit posts we should aggregate for making one training instance. Based on our evaluation on the 2018 dataset, this setting led to very different results. In the official submissions for task 1, we used 1 post per training instance for run 1, 10 posts for run 2, 20 posts for run 3, and 50 posts for run 4. For Task 2, we used 1, 10, 20, 50 and 100 posts for the official runs from 0 to 4.

4.2 Model 2 - Mutual information

Similar to model 1, in this model we attempted to create a training set with positive and negative examples for anorexia and self-harm from Reddit posts. Instead of taking all historical data, we focused on the 1000 new, hot and top posts. Moreover, we used a data-driven approach based on the mutual information measure to extract automatically the relevant positive and negative n-grams for anorexia and self-harm signs [5].

Data A collection containing new, hot and top posts from 50 mental health-related and general subreddits were extracted as candidates for providing positive and negative examples, including all, r/AnorexiaNervosa, r/AskReddit, r/eating_disorders, r/funny, r/movies, r/selfharm, r/sports, r/SuicideWatch, r/television, and r/worldnews. For the anorexia task, the subreddits r/AnorexiaNervosa, r/eating_disorders, r/fatlogic, r/happy and r/progresspics, and, for the self-harm task, r/selfharm, r/SelfHarmScars and r/SuicideWatch subreddits were used as candidates for positive posts, respectively. The other subreddits were used as negative examples for each task.

Training collection and classifier Each post of the positive and negative collections was tokenized, stopword-removed, and stemmed. Furthermore, 1-, 2- and 3-grams were extracted and associated to the respective subreddit. Then, to tag each post the top 200 most relevant n-grams from each collection were used according to their mutual information score. If a post from the positive collection contained more positive n-grams, it was deemed as positive. Similarly, a post from the negative collection was deemed as negative if it contained more negative n-grams. From the 128,170 posts, 5,997 and 88,618 were identified as positive and as negative candidate posts, respectively, for task 1 and 1,279 as positive and 93,944 as negative candidate posts for task 2. The positive and negative candidate posts were then tagged with the mutual information score of the 200 most informative n-grams from the whole collection to create the feature set. This feature set was then used to train a logistic regression and a linear SVM classifiers to categorize posts into anorexia and self-harm categories for tasks 1 and 2, respectively. The SVM and logistic regression classifiers were validated on the 2018 collection, achieving F1-score of 0.63 and 0.73, respectively.

4.3 Model 3 - Convolutional Neural Networks (CNN)

Data Here, we also retrieved a collection of subreddits on anorexia (r/Eating disorders, r/BingEatingDisorder, r/Anorexia, r/AnorexiaNervosa, r/fuckeatingdisorders) and one collection on self-harm (r/selfharm, r/SuicideWatch) and a collection on general topics (r/jokes, r/fitness, r/books, r/teaching, r/writing, r/personal finance) for negative instances to train two CNN models. The subreddit posts were retrieved from a period of one year (2017/11–2018/10). Since there were fewer subreddit posts on anorexia, we sampled 10,000 posts from the general topics as negative instances. This resulted in a corpus of 15,942 posts (positive: 5,942 and negative: 10,000) for anorexia and a corpus of 178,088 posts (positive: 49,845 and negative: 128,243) for self-harm.

Training and classification We first removed all deleted posts and link posts, we then represented the remaining posts using word2vec word embeddings [10] that we trained on a collection of subreddits (200 dimensions) using word2vec CBOW model. We then took a similar approach to that of [4] and applied a convolution operation on a window of 3 and 4 words of the posts, followed by a

Max Pooling layer and a final Sigmoid layer that outputs probability scores.[6] We fixed the sequence length of posts to 300 (shorter input sequences are padded with zeros) and used 10 filters and mini-batch sizes of 50, and used 10% of the data for validation.

4.4 Model 4 - Ensemble model

We further combined the results of three methods based on the weighted normalised score provided by each model. Using the F1-score of the best individual model as reference, the scores of the three models were combined linearly and averaged. If the final score is greater than or equal 0.5, the ensemble model assigns a positive decision to a user.

5 Results and Discussion

5.1 Evaluation metrics

The systems were evaluated based on precision, recall, F1 measure, early risk detection error (*ERDE*), and LatencyTP. ERDE takes into account the correctness of the binary decision and the delay taken by the system to make the decision [7]. While for the first three metrics, the higher the score the better is the system, for *ERDE*, the lower the better. ERDE measure was used with cutoff parameter set to 5 and 50 posts. *LatencyTP* measures the systems delay in detecting positive cases based on the median number of writings. Finally, *Latency-weighted F1* combines the effectiveness of the decision and the delay. Additionally, two other measures of *speed* and *latency-weighted* F-score take into account a penalty based on the median delay for making a positive decision.

5.2 Official results of anorexia risk prediction

We submitted five runs for task 1 and for task 2 using bag-of-word models. Due to some technical issues, we were able to submit the runs for only 11 user posts in task 1 and 8 user posts in task 2. Table 3 and Table 4 present the official results for early prediction of anorexia signs and self-harm risk, respectively. The model using the large Reddit semi-supervised training data with 1 post per training instance (run 1) achieved the best F1 score (0.54) among our submitted models for task 1, and for task 2 (0.46). The results show that combining more than 10 posts of users for training the models decreases the performance drastically. We believe that this is probably related to the characteristics of the task, which simulates an on-the-fly news feed. Hence, processing post streams in chunks jeopardizes the performance of the model.

For task 1, the best document representation was achieved using *tf-idf* weighting scheme with no stopwords removal: personal pronoun words such as *"my"*

[6] We used keras platform, https://keras.io/.

Table 3. Official results of anorexia risk prediction using bag of word models.

	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	*Speed*	*Latency-weighted* F_1
run 0	.42	.07	.12	.09	.08	1	1	.12
run 1	.44	**.70**	**.54**	.06	.03	3	.99	.54
run 2	.73	.11	.19	.08	.08	3	.99	.19
run 3	1	.01	.03	.09	.09	1	1	.03
run 4	0	0	0	-	-	-	-	-

Table 4. Official results of self-harm risk prediction using bag of word models.

	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	*Speed*	*Latency-weighted* F_1
run 0	.52	.41	**.46**	.10	.08	3	.99	.46
run 1	1.0	.05	.09	.12	.11	6.5	.98	.09
run 2	0	0	0	-	-	-	-	-
run 3	0	0	0	-	-	-	-	-
run 4	0	0	0	-	-	-	-	-

Table 5. Statistics on participating runs for anorexia and self-harm and our ranks.

	T1				*T2*			
	P	R	F_1	$ERDE_5$	P	R	F_1	$ERDE_5$
Median	.39	.66	.37	.08	.12	.49	.22	.13
Max	.77	1.0	.71	.17	.71	1.0	.52	.23
Min	0	0	0	.06	0	0	0	.08
Our rank	3	22	11	1	7	11	4	3

weight, *"my" eating*, *"my" boyfriend* seem to be used frequently by people talking about their problems. One effective setting was the use of 2- or 3-grams for detecting collocations, such as *"skipped lunch"*, *"egg white"*, *"light peanut butter"* that are frequently used by users associated with risky behaviors. Table 5 shows the overall statistics for tasks 1 and 2. Our bests models where ranked 11 in terms of F1-score and 1 in terms of ERDE5 for task 1 and ranked 4 in terms of F1-score and 3 in terms of ERDE5 metric for task 2.

5.3 Unofficial evaluation results

We further provide the results of our models described in Section 4 in Table 6 and Table 7 for task 1 and task 2, respectively. From the individual models, Model 1, used in the official run, achieved the best results in terms of F1-score. While Model 3 achieved F1-score of .96 and .98 on the validation sets, its performance on the on-the-fly, post-level test data is limited. Model 3 achieves a recall of 1, however, the precision is quite low due to large number of false positives. This can be explained by the fact that, as mentioned earlier, Model 3 used subreddit posts as a proxy for positive anorexia and self-harm risks during the training phase, but these subreddit collections were different from the actual annotations in the test data. In terms of *ERDE*, Model 1 shows the best performance among all three methods and is quicker in determining the positive cases. Model 2 used a much smaller collection compared to Model 1 ($\mathcal{O}(10^6)$) vs. $\mathcal{O}(10^{12})$), nevertheless, it still performed relatively well for task 1. For task 2, it had almost 50% drop in recall compared to Model 1, jeopardizing the overall F1-score. The ensemble model was created using the linear combination of the best official models and the results of Model 2 and Model 3. As these models are based on very different approaches, we expected to see a significant performance improvement (as seen in [14]). However, as we can see from Table 6 and Table 7, this was not the case, particularly for task 2, for which there was a significant drop in performance for almost all metrics. For task 1, there was some performance gains, such as 4% F1-score (relative); however, this improvement was not extended to the early risk detection error metrics.

Table 6. Additional results on anorexia risk prediction in terms of decision-based metrics against the *official results (best model:run 1).

	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	$Speed$	$Latency\text{-}weighted\ F_1$
*Model 1	.44	.70	.54	.06	.03	3	.99	.54
Model 2	.40	.70	.50	.09	.06	16	.94	.48
Model 3	.09	1	.17	.09	.08	1	1	.17
Ensemble	.46	.73	**.56**	.07	.04	5	.98	.55

Table 7. Additional results on anorexia risk prediction in terms of decision-based metrics against the *official results (best model:run 1).

	P	R	F_1	$ERDE_5$	$ERDE_{50}$	$Latency_{TP}$	$Speed$	$Latency\text{-}weighted\ F_1$
*Model 1	.52	.41	.46	.10	.08	3	.99	.46
Model 2	.47	.22	.30	.12	.12	78	.71	.21
Model 3	.12	1.0	.22	.13	.10	3	.99	.21
Ensemble	.75	.07	.13	.12	.12	13	.95	.13

5.4 Performance change overtime

The performance of the models changes based on the number of posts that they process. Figures 2 and 3 show these changes for both task 1 and task 2, respectively. The performance of Model 1 increases up to around 10 posts for both tasks. The performance of Model 2 increases for about 100 posts for task 1 with the F1-score reaching a maximum at 62% and then it decreases slightly and appears to become stable around 1200 posts with an F1-score of 50%.

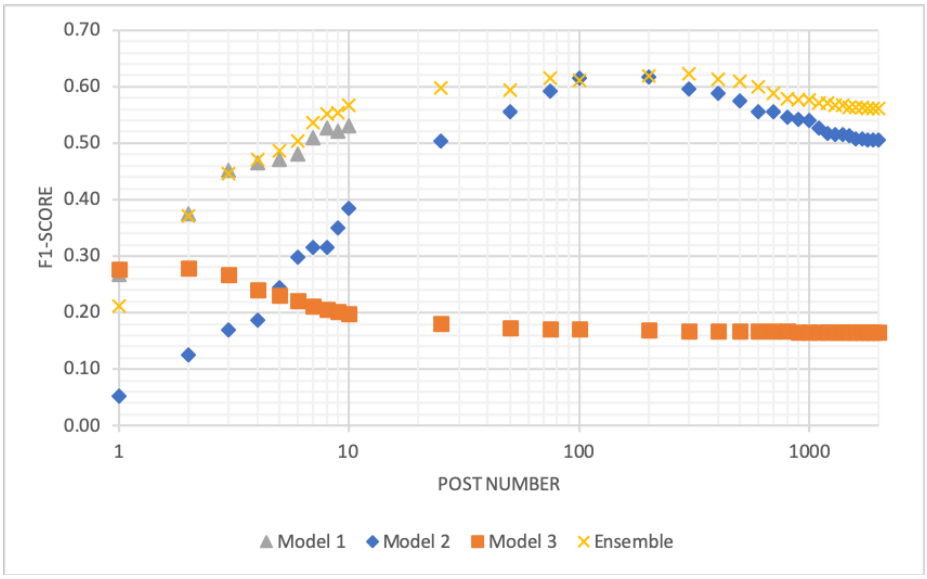

Fig. 2. Model performances based on the number of posts processed for prediction of anorexia. Y-axis is in logarithmic scale.

For task 2, the performance of model 2 increases at about 500 posts up to an F1-score of 32% and then decreases slightly and appears to become stable at around 800 posts with an F1-score of 30%. The performance of model 3 is at its highest (F1-score of 28%) for the first 10 posts and then appears to become stable with an F1 score of 16% for task 1. Similarly for task 2, the performance

of model 3 increases for the first 10 posts and then appears to become stable at F1 score of 21% for the remaining posts. As it is expected for any classifier, as the number of test records increases, the number of false positive also do, hence, justifying the decrease in performance after a certain peak.

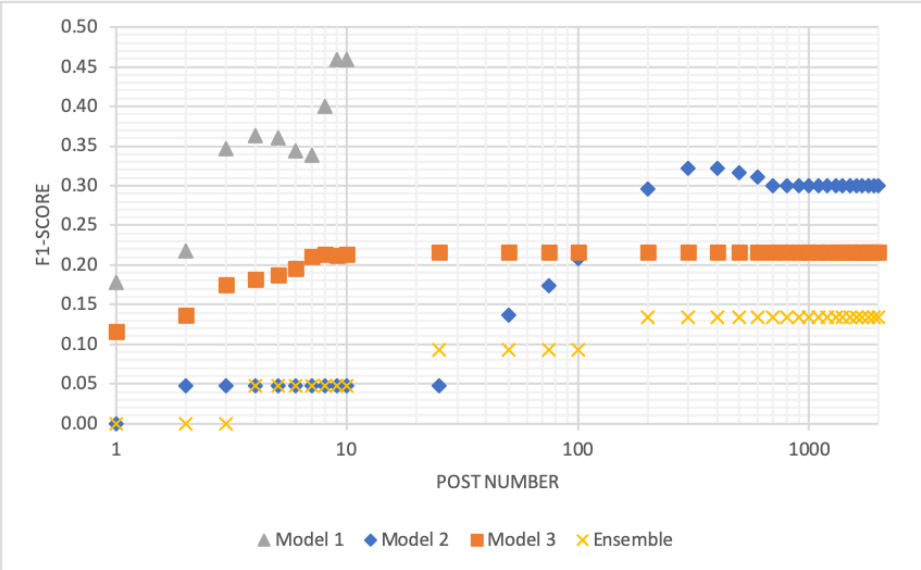

Fig. 3. Model performances based on the number of posts processed for prediction of self-harm. Y-axis is in logarithmic scale.

6 Conclusion

Identifying early signs of mental health disorders among individuals can help early interventions of healthcare systems and lead to better treatment results. In this paper, we presented our data-driven approaches for task 1 and task 2 of the CLEF eRisk 2019 challenge. Our models leverage on existing collections from Reddit for both tasks, without any handcrafted features. Among all the models that we explored, it seems that the use of a very large collection had the most significant impact on the systems performance. Nevertheless, this task proved to be quite challenging and further experiments are needed to understand what features and/or methods are likely to advance the field.

References

1. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diag-

noses. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 1–10 (2015)
2. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych 2015 shared task: Depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 31–39 (2015)
3. De Choudhury, M., De, S.: Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751 (2014)
5. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical review E **69**(6), 066138 (2004)
6. Losada, D.E., Crestani, F., Parapar, J.: CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations. In: CLEF (Working Notes) (2017)
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: Early Risk Prediction on the Internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 343–361. Springer (2018)
8. Losada, D.E., Crestani, F., Parapar, J.: Early Detection of Risks on the Internet: An Exploratory Campaign. In: European Conference on Information Retrieval. pp. 259–266. Springer (2019)
9. Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., Schwartz, H.A.: Clpsych 2018 shared task: Predicting current and future psychological health from childhood essays. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 37–46 (2018)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Milne, D.N., Pink, G., Hachey, B., Calvo, R.A.: CLPsych 2016 shared task: Triaging content in online peer-support forums. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. pp. 118–127 (2016)
12. Niveau, G., Frioud, E., Aguiar, D., Ruch, P., Auckenthaler, O., Baudraz, J., Fracasso, T.: Suicide Notes: Their Utility in Understanding the Motivations Behind Suicide and Preventing Future Ones. Archives of Suicide Research pp. 1–14 (2018)
13. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
14. van Rijen, Paul, a.T.D., Naderi, N., Mottin, L., Knafou, J., Jeffryes, M., Ruch, P.: A data-driven approach for measuring the severity of the signs of depression using reddit posts. In: Proceedings of CLEF (Working Notes) (2019)
15. Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., Pantic, M.: AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. pp. 3–9. ACM (2017)
16. Schwartz, H.A., Eichstaedt, J., Kern, M.L., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.: Towards assessing changes in degree of depression through facebook. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 118–125 (2014)

17. Trotzek, M., Koitka, S., Friedrich, C.M.: Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia. In: CLEF (Working Notes) (2018)
18. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th international workshop on udio/visual emotion challenge. pp. 3–10. ACM (2016)
19. Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M.: AVEC 2014: 3d dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. pp. 3–10. ACM (2014)
20. Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., Pantic, M.: AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. pp. 3–10. ACM (2013)
21. Zirikly, A., Resnik, P., Uzuner, O., Hollingshead, K.: Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. pp. 24–33 (2019)