

BioInfo@UAVR at eRisk 2019: delving into social media texts for the early detection of mental and food disorders

Alina Trifan^[0000-0001-7613-1435] and José Luís Oliveira^[0000-0002-6672-6176]

DETI/IEETA, University of Aveiro, Portugal
{alina.trifan, jlo}@ua.pt

Abstract. This paper describes the participation of the Bioinformatics group of the Institute of Electronics and Engineering Informatics of University of Aveiro in the shared tasks of CLEF eRisk 2019¹. The objective of the eRisk initiative is to encourage research in the area of information retrieval for the automatic detection of risk situations on the internet. The challenge was organized in three tasks, focused on the early detection of anorexia (T1), self-harm (T2) and severity of depression (T3). We addressed these tasks using a mix approach that combines machine learning with psycholinguistics and behavioural patterns. The results obtained validate the use of such patterns in the context of social media mining and motivate future research into this field.

Keywords: information retrieval · early detection · depression · anorexia · psycholinguistic patterns.

1 Introduction

The large volume of written data available through social media attracted the attention of natural language processing researchers over the last years. Social media data has been identified as an emerging opportunity for revolutionizing in-the-moment measures of a broad range of people's thoughts and feelings [13]. Research initiatives such as CLEF Early Risk emerged over the last years as a proof of the importance of this research area. They foster collaborative work on the topic of mental health and social data, and push forward new discoveries and insights. As a practical outcome that the eRisk initiative encourages is the fact that triaging online social networks data or public forums enables the identification of content that requires the attention of moderators to ensure that urgent content can be responded to more quickly and consistently. Over the last years, the focus of these shared tasks was the early identification of people susceptible to depression or suffering from food disorders.

¹ <http://early.irlab.org/>

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

Prevention and early identification of mental and food disorders by means that are complimentary to traditional medical approaches have the ability to mitigate the under-supply of mental health facilities by advancing different types of counseling or support for the ones in need, such as connecting a depressed person to resources or peer support when they most need it [11]. Using social data has yet another advantage with respect to the stigma associated to mental health screening, as it can lead to treatment of people who are otherwise less inclined to pursue clinical services [8]. Such approaches can provide new opportunities for early detection and intervention and they have the potential to open new insights on research of the causes and mechanisms of mental health [4]. Two of the tasks proposed by the CLEF eRisk 2019 initiative focus on the early detection of signs of anorexia and signs of self-harm, respectively. For this purpose, social media posts had to be sequentially processed and a decision should have been emitted as soon as possible. The classification metrics used in these tasks take into consideration the delay in emitting a positive classification of an user suffering of self-harm ideation or anorexia. The third task of this year’s challenge was aimed at estimating the level of depression from a thread of user submissions.

This paper describes the participation of the BioInfo@UAVR team in the CLEF eRisk 2019 tasks. In our approach, we combined standard machine learning algorithms with psycholinguistics and behavioral patterns derived from the literature. The methodology and associate results are presented in this paper, along with proposed future work. The rest of this paper is organized as follows: section 2 outlines the research background behind the proposed tasks. The following three sections are dedicated to the description of the tasks, and include both the methodologies used and the results obtained. We conclude the paper and discuss future work in section 6.

2 Background

The widespread use of social media, combined with the rapid development of computational infrastructures to support big data and the maturation of natural language processing and machine learning technologies, offer exciting possibilities for the improvement of both population-level and individual-level health [10]. The Internet and social media have quickly become major sources of health information, providing both broad and targeted exposure to such information as well as facilitating information-seeking and sharing. As people increasingly turn to social media for news and information, these platforms can serve as novel sources of observational data for infodemiology, public health surveillance, tracking health attitudes and behavioral intention, and measuring community-level psychological characteristics related to health outcomes [26, 20, 14, 29, 17, 6]. Patients with chronic health conditions use online health communities to seek support and information to help manage their condition. The automatic mining of forum posts can provide help assisting patients in need of clinical expertise by getting proper health [25]. Moreover, patients can realize what are the feelings or opinions of users who have similar conditions, and caregivers may

better understand how users’ feelings differ under various conditions and then provide proper healthcare for their patients [27].

Sentiment analysis has been applied to social media to identify important public health issues, such as public attitudes towards vaccination or towards marijuana, just to name a few examples. Emotion tweets can be utilized to detect and monitor disease outbreaks, which suggests that emotion classification could help distinguish outbreak-related tweets from other disease discussion [18, 19, 7]. This social data mining can improve our understanding of the determinants and consequences of well-being, which is correlated with outcomes of both mental and physical health [22].

3 Task 1 - Early detection of signs of anorexia

Task 1 consisted in sequentially processing pieces of evidence and detect early traces of anorexia, as soon as possible. The collection contains writings of social media content from two categories of users: anorexia and non-anorexia. A labelled training collection was released prior to the evaluation period. For the test stage a server that iteratively releases user writings was set up by the organization. After each round of writings release a decision had to be emitted. Classifying a user as suffering from anorexia was considered an irreversible decision, while a decision of non-anorexia was open to updates in the following rounds of decisions.

3.1 Dataset description

The training and test collection for this task have the same format as the collection described in [15]. The source of data is the same as for previous eRisk challenges, namely eRisk 2017 and 2018. They represent collections of writings (posts or comments) from a set of social media users and, for each user, the collection contains a sequence of writings in chronological order. The characteristics of the training set are presented in Table 1.

Table 1. Task1 training dataset.

	Anorexia	Non-anorexia
#subjects	61	411
#posts	24874	228878
avg #posts/subject	398.75	566.2
avg #words/post	39.95	20.9

3.2 Metrics

The evaluation metrics that have been regularly used for the eRisk challenges is ERDE, the early risk detection measure proposed by Losada et al. [15]. As

identified in this year’s overview report [16], this measure has several drawbacks, which led to the inclusion of alternative evaluation metrics. As such, $F_{latency}$ a measure proposed by Sadeque et al. [24] was also used. This measure takes into consideration the effectiveness of the decision (estimated with the F measure) and the delay for emitting the decision. A perfect system would get an $F_{latency}$ of 1. These metrics are further complemented with a ranking evaluation of the systems after seeing k writings, with varying k.

3.3 Methods

In the preprocessing step of our approach the posts are lowercased and tokenized, after removing all non-alphabetic characters. Stopwords are filtered, based on the stopwords list of the Natural Language Toolkit². We explored both incremental and online training with the following three classifiers: Multinomial Naive Bayes, linear Support Vector Machine with Stochastic Gradient Descent and Passive Aggressive. For the out of core classification, we trained the classifiers with batches of 500 users data. The batch size is not expected to have an impact on the performance of the classifiers³. For each of these classifiers, we performed a grid search over the validation dataset in order to identify the best parameters that characterize them. We considered Bag of Words features for the three classifiers and we applied counts and tf-idf based feature weighting. The classifier that led to better results on the validation corpus was the SVM with SGD classifier, with a stopping criterion of 1e-3 and a modified Huber loss.

The number of writings per user was not known in the test stage. Our strategy for early detection was to only delay emitting a positive decision during the first 3 rounds of getting server writings. This would allow us to get at least 3 writings for each user, without compromising too much the delay in the response. Another important aspect of our submission was the fact that we processed each thread of user writings in real time. This means we did not use any offline knowledge or processing and we provided a response as fast as possible after getting a round of user writings from the server.

3.4 Results

The results obtained are shown in Table 2, along with the best results in this task, for comparison. The results of all participating teams can be found in [16]. Compared to the remaining 12 participant teams, our team was the only one to submit only one run of results. Most teams used five runs, which was the maximum number of runs allowed. Our results place us in the middle of the team rankings for this task.

In terms of ranking, after processing 1, 100, 500 and 1000 writings, we obtained constant values for P@10 (0.6), NDCG@10 (0.59) and NDCG@100 (0.47).

² <https://www.nltk.org/>

³ <http://scikit-learn.org/0.19/modules/scaling-strategies.html>

Table 2. Evaluation of BioInfo@UAVR’s submission in Task 1. The best results were added for comparison.

	P	R	F1	ERDE5	ERDE50	latency	speed	latency-weighted F1
BioInfo@UAVR	.32	.44	.37	.06	.06	1	1	.37
Best results	.64	.79	.71	.06	.03	7	.98	.69

4 Task 2 - Early detection of signs of self-harm

This task considered the early detection of users of social media prone to self-harm themselves. As no training dataset was provided, we approached this task as a cross-validation task rather than an unsupervised classification. Self-harm ideation often relates to depression and poor mental health, therefore we were interested in understanding how a classifier trained on a depression corpus of social media writing would perform in the test stage.

4.1 Dataset description

The training dataset used in this task is the one proposed by Yates et al. [28], publicly available through a signed user agreement that emphasises on data protection and proper acknowledgements. The dataset consists of all Reddit users who made a post between January and October 2016, matching high-precision patterns of self-reported diagnosis (e.g. “I was diagnosed with depression”). The depressed users were matched by control users, who have never posted in a subreddit related to mental health and never used a term related to it. In order to avoid a straight-forward separation of the two groups, all posts of diagnosed users related to depression or mental health were removed. In the end, 9210 diagnosed users were matched by 107 274 control users. Each user in the dataset has an average of 969 posts (median 646) and the mean post length is 148 tokens (median 74).

4.2 Metrics

The metrics used for the evaluation of this tasks’s submission are identical to the ones used in Task 1.

4.3 Methods

For this task we followed a standard processing stream for text classification. We initially split the dataset into training and validation chunks, with a ratio of 2:1. We considered Bag of Words (BoW) and tf-idf based feature weighting with linear Support Vector Machine with Stochastic Gradient Descent and Passive Aggressive classifiers. We trained and validated both classifiers on the validation corpus.

The SVM led to slightly better results in terms of F1 in the validation stage, so we retrained the model with the whole corpus (training + validation). In

the competition’s test stage we used this trained model to predict the class of self-harm or no self-harm of the user’s writings provided by the iterative server. Our strategy for this task was very similar to the one in Task 1. We only started emitting decisions in the forth round of server writings and we did all the classification online, without applying any offline knowledge.

4.4 Results

The results obtained are shown in Table 3, along with the best results in this task, for comparison. Our approach ranked 4th both in terms of F1 and latency-weighted F1 in a total of 33 submissions of 8 different teams. The results of all participating team can be found in [16].

Table 3. Evaluation of BioInfo@UAVR’s submission in Task 2. The best results for each metric were added for comparison.

	P	R	F1	ERDE5	ERDE50	latency	speed	latency-weighted F1
BioInfo@UAVR	.55	.39	.46	.11	.08	6	.98	.45
Best results	.71	.41	.52	.09	.07	2	1	.52

These results stand as a proof of the links between depression and self-harm and are of a particular importance as the training dataset was completely different from the test one. This task was open to algorithmic imagination during the training stage as no training data was provided, along with the no disclosure of any information about the test dataset prior to the test stage. The training dataset that we used was agnostic to the structure or type of data that was later released in the test stage. Our team processed the test data online, meaning no external knowledge or processing was performed after having access to the first round of test submissions.

5 Task 3 - Estimating the level of depression

This task was aimed at exploring the viability of automatically estimating the severity of multiple symptoms associated with depression [16]. Given the users history of writings, participants had to work out a solution for predicting the users response to each individual question included in Beck’s Depression Inventory Questionnaire (BDI) [5]. The questionnaire assesses the presence of feelings like sadness, pessimism, loss of energy, hunger/loss of appetite, etc. For each individual question, a numeric value between 0 and 3 is considered a valid answer, with the exception of two questions, whose possible answers were: 0, 1a, 1b, 2a, 2b, 3a or 3b.

5.1 Dataset description

A dataset with 20 files, one file per user was provided. Each file contained the history of writings of the respective user. The number of writings per user varied from 30 to 1511. The average number of writings of the dataset was 548, with a median of 328.5.

5.2 Metrics

The ground truth used for the evaluation of the responses provided by the participants in this task were the questionnaires filled in by the social media users whose writings were provided in the dataset. For each user of the dataset the respective writings were extracted right after having provided the filled questionnaire.

The evaluation metrics reflected the differences between the answers of the questionnaire provided by the task participants and the ones provided by the users that were part of the dataset. Moreover, in the psychological domain it is customary to associate depression levels with categories. Depression levels are defined as the sum of all answers of the 21 questions of the questionnaire. The following depression categories were used for further extension of the evaluation metrics:

- minimal depression - [0..9]
- mild depression - [10..18]
- moderate depression - [19..29]
- severe depression - [30..63]

The following metrics were considered for the evaluation of the results [16]:

- *Hit Rate (HR)* - the ratio of cases where the automatic questionnaire has exactly the same answer as the real questionnaire.
- *Average Hit Rate (AHR)* - HR averaged over all users.
- *Closeness Rate (CR)* - the absolute difference between the real and the participant provided answer.
- *Average Closeness Rate (ACR)* - CR averaged over all users.
- *Difference between overall depression levels (DODL)*.
- *Average DODL (ADODL)* - DODL averaged over all users.
- *Depression Category Hit Rate (DCHR)* - the fraction of cases where the automated questionnaire led to a depression category that is equivalent to the depression category obtained from the real questionnaire.

5.3 Methods

Our approach for solving this task was a rule-based one and each rule was modelled with reference to several behavioral and psycholinguistics patterns that are known to be associated with the state of depression (Table 4). The reduced size of the dataset, in terms of users and the small number of writings per user, along

with the lack of a training set or any ground truth weight, led to the choice of a rule-based approach rather than the use of standard machine learning algorithms.

Moreover, we explored the correlation between some of the questions by dividing the 21 questions into 6 groups. All questions belonging to a given group were scored with the same answer or numeric value. The 6 groups and the included categories (or question names) were:

1. Depression - suicidal thoughts, pessimism, past failure, self dislike, sadness, loss of pleasure, loss of interest and loss of sex
2. Guilt - guilty and punishment feelings, self criticalness, crying, worthlessness
3. Appetite - changes in appetite
4. Anxiety - agitation, indecisiveness, irritability
5. Fatigue - tiredness, loss of energy, concentration difficulty
6. Sleep - sleeping patterns

Table 4 overviews the textual and behavioral patterns modelled for each of the 6 groups. For each category, a score was calculated for each user as a normalized value of the number of occurrences of the features considered for each category with respect to the total number of occurrences of the same features over the dataset. These scores were then normalized to the interval $[0,3]$ based on pre-defined thresholds extracted from the histograms of occurrences. An example of such histogram and the threshold derived from it are shown in Fig. 1. In this example, a depression score lower than 0.3 would be converted to a 0, a score in the range of $(0.3, 0.5)$ leads to a 1, a score equal or higher than 0.5 but lower than 1 represents a final score of 2 and anything over 1 means the final answer for the questions in the depression category will be 3.

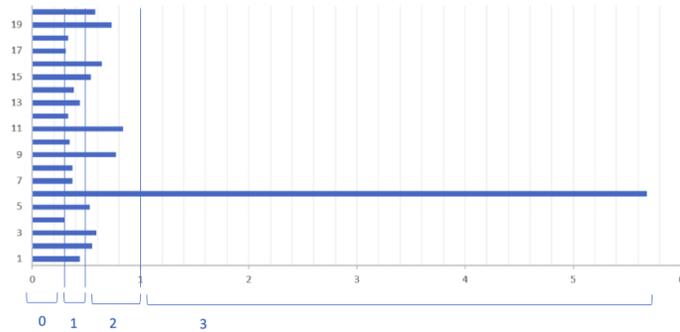


Fig. 1. Histogram of the depression scores calculated for each of the 20 users of the dataset. The vertical bold lines represent the threshold values for the normalization of the scores to the integer values defined as possible answers. In this example user 6 stands out as her text history led to a much higher depression score than the average. This is the case of what seems to be a support pal - this particular user employed extensive depression related vocabulary to provide support and comfort.

Table 4. Details of the textual features considered for each category score.

Depression	<p>Lexical category of a user’s text - depressed users tend to have an overall more negative connotation of their texts [21, 12]. To this purpose we employed the TextBlob library [1] in order to calculate the average polarity of a user’s writings.</p> <p>Use of self-related words (e.g: I, myself, mine) - depressed users tend to use them more often in their writings [9, 23]</p> <p>Use of absolutist words - Al-Mosaiwi et al. [3] recently showed that anxiety, depression, and suicidal ideation forums contained more absolutist words than control forums. The list of absolutist words used is presented next in Table 5.</p> <p>Referrals to any of the anti-depressants listed by WebMD [2].</p> <p>Mentions of words related to mental disorders, (e.g.:depression, bipolar, schizophrenia, psychotic, ocd).</p> <p>Writings timestamps - depressed users tend to write more at late hours of the night.</p>
Guilt	Use of the words cry, guilt and their derivatives.
Appetite	Use of the words hunger, appetite, eat, food and their derivatives.
Anxiety	Use of the words sleep, anxious and their derivatives.
	Writings timestamps.
Fatigue	Use of the words irritated, fatigue, tired and their derivatives.
Sleep	Same as for fatigue, along with the writing timestamps.

Table 5. Absolutist words validated by Al-Mosaiwi et al. [3].

absolutely	all	always	complete	completely
constant	constantly	definitely	entire	ever
every	everyone	everything	full	must
never	nothing	total	whole	

5.4 Results

Task participants had to provide a result file containing 20 lines, one for each user in the dataset. Each line contained the username and 21 values that corresponded to the answers of the 21 questions included in Beck’s Depression Inventory. The results obtained by our team are presented in Table 6, along with the best results obtained in this task, for each of the metrics. The results of all participating teams can be found in [16]. In this task 8 different teams submitted 18 runs and no single team was able to achieve best results for each of the metrics. Overall, this task’s results were quite homogeneous with little variations from team to team. This can be seen as an indication of both the difficulty of the task and possibly the similarity in the approaches adopted by the participating teams.

Table 6. Evaluation of BioInfo@UAVR’s submission in Task 3. The best results for each metric were added for comparison. To note that no single team achieved the best results for all metrics.

	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	34.05%	66.43%	77.70%	25.00%
Best scores	41.43%	71.27%	81.03%	45.00%

6 Conclusions and Future Work

We presented in this paper the results of our team’s participation in the eRisk2019 shared tasks. Through this challenge we came to understand that the analysis of social media texts has the potential to provide insights into understanding a user’s mental health status and for the early detection of possible related diseases. Being this our first participation in these challenges, we understand there is still room to improve our methodologies. Nevertheless, the results obtained encourage us to further contribute to this area of research.

As future work we plan to combine the methodologies used in the first two tasks with the one of task 3. We believe the results obtained in the first two tasks can be further improved by the use of psycholinguistic patterns that relate to self-harm ideation or anorexia. With respect to task 3, we are keen in understanding how a classifier trained on a depression or self-harm corpus would perform on scoring the level of depression for this task.

Acknowledgments

This work was supported by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

References

1. Textblob (2019), <https://textblob.readthedocs.io/en/dev/>
2. WebMD (2019), <https://www.webmd.com/depression/guide/depression-medications-antidepressants>
3. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science* p. 2167702617747074 (2018)
4. Arseniev-Koehler, A., Mozgai, S., Scherer, S.: What type of happiness are you looking for?-A closer look at detecting mental health from language. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. pp. 1–12 (2018)
5. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. *Archives of general psychiatry* **4**(6), 561–571 (1961)
6. Benton, A., Coppersmith, G., Dredze, M.: Ethical research protocols for social media health research. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. pp. 94–102 (2017)
7. Bravo-Marquez, F., Frank, E., Mohammad, S.M., Pfahringer, B.: Determining word-emotion associations from tweets by multi-label classification. In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. pp. 536–539. IEEE (2016)
8. Bruffaerts, R., Mortier, P., Kiekens, G., Auerbach, R.P., Cuijpers, P., Demeyttenaere, K., Green, J.G., Nock, M.K., Kessler, R.C.: Mental health problems in college freshmen: Prevalence and academic functioning. *Journal of affective disorders* **225**, 97–103 (2018)
9. Chung, C., Pennebaker, J.W.: The psychological functions of function words. *Social communication* **1**, 343–359 (2007)
10. Conway, M., OConnor, D.: Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology* **9**, 77–82 (2016)
11. Coppersmith, G., Leary, R., Whyne, E., Wood, T.: Quantifying suicidal ideation via language usage on social media. In: *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM* (2015)
12. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. *ICWSM* **13**, 1–10 (2013)
13. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* **18**, 43–49 (2017)
14. Kim, Y., Huang, J., Emery, S.: Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of medical Internet research* **18**(2) (2016)
15. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. pp. 28–39. Springer (2016)
16. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019*. Springer International Publishing, Lugano, Switzerland (2019)
17. Loveys, K., Crutchley, P., Wyatt, E., Coppersmith, G.: Small but mighty: Affective micropatterns for quantifying mental health from social media language. In:

- Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality. pp. 85–95 (2017)
18. Mohammad, S.M., Bravo-Marquez, F.: Emotion intensities in tweets. arXiv preprint arXiv:1708.03696 (2017)
 19. Mohammad, S.M., Kiritchenko, S.: Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence* **31**(2), 301–326 (2015)
 20. Mollema, L., Harmsen, I.A., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., Kok, G., Ruiters, R., Das, E.: Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. *Journal of medical Internet research* **17**(5) (2015)
 21. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in twitter. In: *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*. vol. 2012, pp. 1–8. ACM New York, NY (2012)
 22. Paul, M.J., Sarker, A., Brownstein, J.S., Nikfarjam, A., Scotch, M., Smith, K.L., Gonzalez, G.: Social media mining for public health monitoring and surveillance. In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. pp. 468–479. World Scientific (2016)
 23. Rude, S., Gortner, E.M., Pennebaker, J.: Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* **18**(8), 1121–1133 (2004)
 24. Sadeque, F., Xu, D., Bethard, S.: Measuring the latency of depression detection in social media. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. pp. 495–503. ACM (2018)
 25. VanDam, C., Kanthawala, S., Pratt, W., Chai, J., Huh, J.: Detecting clinically related content in online patient posts. *Journal of Biomedical Informatics* **75**, 96–106 (2017)
 26. Vaterlaus, J.M., Patten, E.V., Roche, C., Young, J.A.: # gettinghealthy: The perceived influence of social media on young adult health behaviors. *Computers in Human Behavior* **45**, 151–157 (2015)
 27. Yang, F.C., Lee, A.J., Kuo, S.C.: Mining health social media with sentiment analysis. *Journal of medical systems* **40**(11), 236 (2016)
 28. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. p. 29682978. Association for Computational Linguistics (2017)
 29. Zhang, J., Brackbill, D., Yang, S., Centola, D.: Identifying the effects of social media on health behavior: Data from a large-scale online experiment. *Data in brief* **5**, 453–457 (2015)