UPF's Participation at the CLEF eRisk 2018: Early Risk Prediction on the Internet

Diana Ramírez-Cifuentes and Ana Freire

Web Science and Social Computing Research Group
Universitat Pompeu Fabra, Barcelona
Carrer Tanger, 122-140, 08018 , Barcelona, Spain
{diana.ramirez,ana.freire}@upf.edu

Abstract. This paper describes the participation of the Web Science and Social Computing Research Group from the Universitat Pompeu Fabra, Barcelona (UPF) at CLEF 2018 eRisk Lab[1]. Its main goal, divided in two different tasks, is to detect, with enough anticipation, cases of depression (T1) and anorexia (T2) given a labeled dataset with texts written by social media users. Identifying depressed and anorexic individuals by using automatic early detection methods, can provide experts a tool to do further research regarding these conditions, and help people living with them. Our proposal presents several machine learning models that rely on features based on linguistic information, domain-specific vocabulary and psychological processes. The results, regarding the F-Score, place our best models among the top 5 approaches for both tasks.

Keywords: early risk detection · social media · depression · anorexia · machine learning

1 Introduction

Symptoms associated with mental illnesses can be observable on online social networks and web forums [19]. Automated methods have been developed in order to detect depression and other mental illnesses by analysing user-generated data in social media, as stated in the review by Guntuku's et al. [13]. These methods usually rely on classification algorithms that do not consider the delay in detecting positive cases. Losada et al. [16] proposed a temporal-aware risk detection benchmark in order to consider, not only the accuracy of the decisions taken by the algorithms, but also the temporal dimension.

The early detection of signs of depression and anorexia tasks, as part of the CLEF eRisk 2018 challenge, consisted in sequentially processing texts posted by users in social media, and detecting traces of depression or anorexia as early as possible [16, 18]. The texts were meant to be processed in the order they were created, for a further capability of the system to analyse the interaction between users in online media, mostly blogs and social networks, in real time.

[1] http://early.irlab.org/

UPF submitted the results obtained by 4 different models for each task. The process for obtaining them, and the results of their application are described in this paper, along with the proposed improvements and possible further work on the subject. The remainder of this paper is structured as follows: Section 2 reports the related work in detecting mental illnesses in social media and the application of early risk measures. Section 3 describes the two tasks addressed. Section 4 shows our research proposal, focusing on the feature extraction process and the learning algorithms used for both tasks. We report our experimental setup in Section 5, followed by our results and findings in Section 6. Finally, Section 7 summarises our conclusions.

2 Related Work

Studies have been conducted on how the usage of social media sites is correlated with mental illnesses in users [13,25]. The language and words, expressed by users in social media texts, may indicate feelings of self-hatred, guilt, helplessness, and worthlessness, which are all elements that characterize depression [11]. People with eating disorders, such as anorexia and bulimia, can be identified by the usage of certain keywords that characterize and promote these conditions [3,28].

Predictive models are built to perform the automated analysis of social media. These models use features or variables that have been extracted from labeled user-generated data [13]. To collect the data, participants are either recruited to take a survey and share their social network account data [11, 23, 27], or data is collected from public online sources like Twitter, Facebook or Reddit [3–5, 9, 14, 22, 28].

Regarding the features that are extracted to build predictive models, the most common ones are those related to the users' texts such as: frequencies of each word or multiple words (N-grams) [24, 27], topics [21, 24, 27], and features obtained using dictionaries like LIWC[2] to measure the usage of self references, social words and emotions [10, 11]. Features based on sentiment analysis are also obtained by calculating the subjectivity or polarity of a phrase [10, 11, 27]. Among other features used, we can find the ones obtained by the analysis of the user activity, like the posting frequency in different periods of the day and year [8, 10]. There are also some research works that obtain features from the relationships between users, considering the number of friends, or followers [11].

As part of the related work, some studies address the importance of an early detection of depression signs, and do an analysis with data prior to the diagnosis [11, 27]. However, the work of Losada et al. [17] proposes a new metric to measure the effectiveness of early alert systems and presents a method for detecting early traces of depression. This provides a measure to compare early detection algorithms in a systematic way.

[2] http://liwc.wpengine.com/.

3 Tasks Description

T1 and T2 consisted in analysing a collection composed by chronologically ordered writings (posts or comments) from a set of social media users [18]. For T1, users were labeled as depressed and non-depressed, and for T2, users were labeled as anorexic and non-anorexic. The collection of writings of each user was split into 10 chunks, with a 10% of the total stored messages of the user in each chunk. There were two stages for each task: a train stage for which the whole history of writings for a set of training users was provided, along with the ground truth; and a test stage, which consisted of 10 sequential releases of data corresponding to each of the 10 chunks with the writings of the test users. Results were meant to be submitted after each release with either a decision for a user, e.g. depressed or not depressed, or no decision, meaning that the system required to see more chunks before deciding. This choice had to be made for each user in the collection. The metrics used for the evaluation of the system were Precision, Recall, F1 Score, and the Early Risk Detection Error (ERDE) metric, proposed in [16], which penalizes the delay in detecting positive cases.

4 Proposal

The proposal defined for the challenge, which is common for tasks 1 and 2, combines a set of features extracted from the concatenated writings of each user. With these features, a model is trained to be applied afterwards to process the users' test text streams, for each task's dataset. To process the writings, the *dynamic method* proposed in [16] is used. This method consisted in building incrementally a representation of each user, and then applying a classifier, which was previously trained with all the users writings. Following this approach, a decision is made if the classifier outputs a confidence value above a given threshold.

4.1 Feature Extraction

The features we considered aim to characterise the content of the writings and provide statistical measures based on the posting frequency and length of the texts. The details on these features are explained below, and a summary can be found in Table 1.

Linguistic, psychological processes and depression related vocabulary: the content of the users' writings was characterized by calculating some scores. This was done based on the frequency of words belonging to the categories of the LIWC2007 dictionary [20], which has been previously used in detecting mental health issues [7, 8]. Scores based on linguistic and psychological processes, as well as personal concerns and spoken categories were obtained. The scores were calculated normalising the frequencies of words by the total number of words in the writings of a user. Given that certain words could belong to multiple categories, the normalization value was augmented in one each time a word was part of more than one category. A feature value was calculated for each of the categories defined in the LIWC2007 dictionary, the list and description of these categories can be found in [20].

For T1, two additional domain-based features were obtained by defining antidepressants and absolutist words categories. In this sense, a list of the 10 leading psychiatric drugs as published in [26], and a set of absolutist words based on the work of [2], were added. This last study concluded that the elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation.

For T2, in addition to the LIWC ones, 9 more features were defined by creating categories of words that belonged to domains related with anorexia. The vocabulary for these categories was obtained from the codebook's domains and sample keywords defined in [3]. The domains are: anorexia, body image, body weight, food and meals, eating, caloric restriction, binge, compensatory behavior, and exercises. Each domain was defined by a list of keywords as stated in [3].

N-grams: extensively used in text mining and natural language processing tasks [12]. They consist of sets of co-occuring words within a given window (N). Some previous works have considered them as features for detecting depression and eating disorders [24, 27]. For the implemented approaches, a $tf \cdot idf$ vectorization was done from the unigrams, bigrams and trigrams of the training set writings. For this step, the *TfIdfVectorizer* from the *scikit-learn* Python library[3] was used, with a stop-words list and the removal of the n-grams that appeared in less than 20 documents. The content of a document was defined by the concatenation of all the writings of a user from all the chunks, in the training phase. Trigrams did not improve the results and therefore they were not used in the models delivered, as stated in 1.

Statistical features: The number of writings of each user and the median of the words used per post were defined as features. For the models sent, these features were discarded given that they seemed to provide good results at the training stage, but the expected results were not obtained when they were tested with the *dynamic method*. Since these features were excluded from the models delivered, they are described as not used in Table 1

Feature with weighted scores: For T1, an additional feature was defined by adding the weighted values of certain features obtained from the LIWC2007 dictionary categories. The features were selected based on the top 4 LIWC categories that were strongly correlated with positive depression cases, as stated in [2]. The antidepressants and absolutist words categories were considered as well. The words belonging to these categories were given a higher weight if they

[3] http://scikit-learn.org/

were found in a subject's writing. Either the same or different weights could be assigned to each category.

In the same way, for T2, a feature was obtained based on the combination of the weighted values of the 9 features based on the categories of words that belonged to domains related with anorexia.

Table 1. Features considered for T1 and T2 in the models delivered.

Feature Type	**Details and resources**	**Used in T1**	**Used in T2**
Linguistic and psychological processes, and depression vocabulary	Depression vocabulary	yes	no
	Anorexia vocabulary	no	yes
	LIWC	yes	yes
N-grams	unigrams	yes	yes
	bigrams	yes	yes
	trigrams	no	no
Statistical features	number of posts per user	no	no
	median number of words per post	no	no
Features with added weighted scores	Addition of the weighted scores of depression related features	yes	no
	Addition of the weighted scores of anorexia related features	no	yes

4.2 Learning Algorithms

Two prediction methods were explored, i.e., Logistic Regression and Random Forest, as they have been used previously as classifiers for similar tasks [16,21] They are briefly explained bellow:

Logistic Regression: statistical method used to predict a binary outcome given a set of independent variables. It predicts the probability of occurrence of an event by fitting data to a logit function [1].

Random Forest: This classification method works by building many decision trees at training time. For the classification tasks, it outputs the class that is the mode of the classes of the individual trees [6].

5 Experimental Setup

The main objective of our proposed models is to detect the highest amount of positive cases, and to do it as soon as possible, minimizing the ERDE and maximizing the F1 Score. *Python 3.6.5*[4] and, in particular, the *scikit-learn Python* library was used for the implementation of the proposed methods.

Using the training data provided for T1, we applied 10-fold cross validation and optimized the parameters through grid search in order to maximize the F1 Score. Each instance of this dataset was defined by the features mentioned in section 4.1 and represented one user. For each user, the features were extracted from the sequentially-concatenated writings of all their chunks. The provided test set allowed us to evaluate the behavior of the dynamic method. Also, this set was used to define a threshold (see Table 2) that represents the minimum probability value required by an instance to be classified as positive. The definition of this threshold contributed to the minimization of the ERDE. The performance of the method was evaluated in terms of the evaluation measures.

Similarly for T2, with all the training data provided we chose to do a 10-fold cross validation combined with grid search in order to optimize the parameters of the algorithms used. The models obtained were afterwards used to process the writings of the test data, applying the dynamic method.

Table 2. Description of the four models designed

	T1		T2	
Model	Features	Configuration	Features	Configuration
UPFA	LIWC: 64 features	Linear Regression	LIWC: 64 features	Linear Regression
	unigrams:12655 features	Threshold=0.75	unigrams: 4303 features	Threshold=0.75
UPFB	LIWC: 64 features	Random Forest	LIWC: 64 features	Random Forest
	unigrams:12655 features	Threshold=0.5	unigrams: 4303 features	Threshold=0.5
UPFC	LIWC: 64 features	Linear Regression	LIWC: 64 features	Linear Regression
	Unigrams and bigrams:18006 features	Threshold=0.75	Unigrams and bigrams: 4970 features	Threshold=0.75
	Depression vocabulary: 2 features		Anorexia vocabulary: 9 features	
	Feature with depression weighted scores: 1 feature		Feature with depression weighted scores: 1 feature	
UPFD	LIWC: 64 features	Random Forest	LIWC: 64 features	Random Forest
	Unigrams and bigrams:18006 features	Threshold=0.55	Unigrams and bigrams: 4970 features	Threshold=0.55
	Depression vocabulary: 2 features		Anorexia vocabulary: 9 features	
	Feature with depression weighted scores: 1 feature		Feature with depression weighted scores: 1 feature	

6 Results

We designed four different models for each task, named *UPFA, UPFB, UPFC* and *UPFD*. Each model contained a particular set of features, and was created by applying either Logistic Regression or Random Forest classifiers. During the

[4] https://docs.python.org/3/

training phase for both tasks, the usage of the statistical features did not improve the results of the prediction with the test set, despite having offered promising results with the training set. Trigrams did not offer better results either. Therefore, these features were excluded from the selected models sent to the challenge.

The details of each model are described in Table 2. The scores for all the models sent for T1 and T2 can be found in [18]. Note that, regarding the ERDE, the values obtained do not evaluate the the actual performance of the models, as the decision files were not delivered to the challenge right from the first chunk and therefore the lab evaluation had to assume that the decision of our classifiers for the earlier chunks was *no decision* for all subjects. The results obtained with *UPFA* and *UPFB* were delivered after the release of the eighth chunk since our team engaged into the lab tasks at an advanced stage. The results of models *UPFC* and *UPFD* were delivered only after the tenth chunk was released, due to a late refinement of these models. We detail the results obtained for each task in the following sections.

6.1 Task 1: Depression

In this task, the best F1 score value (0.55) was provided by *UPFA* model, based on Logistic Regression, the use of unigrams and the categories classified by LIWC. Also, the best ERDE{5,50} scores were reported by the same model. Table 3 also reports Precision, Recall and ERDE scores for *UPFA*. Regarding ERDE, Table 3 reports six different ERDE measures, organised in 3 subsets:

- ERDE{5,50} challenge: the scores reported by the challenge organisers, considering a late delivery of our results.
- ERDE{5,50} chunks: the scores calculated assuming that the results were sent on time for all the chunks.
- ERDE{5,50} writings: the scores calculated with the exact number of writings that were analyzed by each model before making a decision.

The results show that processing the streams dynamically, writing per writing, instead of chunk by chunk, reduces the ERDE value. Also, the Logistic Regression classifiers provided better results compared to the models where Random Forest was applied.

Table 4 reports the results after processing each chunk with the *dynamic method*. Focusing on T1, as more chunks are analyzed, the F1 score increases, and so the precision and recall. The ERDE decreases after analysing the second chunk, and starts to slightly increase afterwards. Regarding ERDE50, the percentage mostly decreases after processing a new chunk. With all chunks processed, we found that the system got the highest amount of true positive cases (47%), right after processing the first chunk, but this is precisely when the highest amount of false positive cases are predicted too (76%).

Based on the F1 Score, our best model got the fifth place among the 45 models presented for T1 [18]. Considering the *ERDE writings* score we can see that our model would have obtained the lowest value for the ERDE50 measure among all the lab results, with a percentage of 6.41%. In this case, we assume that the other models needed to see all the writings of the last chunk from which they made a decision.

6.2 Task 2: Anorexia

The best model for T2 was *UPFC* with a F1 Score of 0,73. Regarding ERDE score, *UPFD* reported the best score for ERDE5 (12.93%) and *UPFA* for ERDE50 (11.34%). As in T1, Table 3 displays the ERDE chunks and ERDE writings. We can see that processing the streams writing per writing and Logistic Regression classifiers provided better results.

From Table 4 we observe that, even though the recall increases considerably after processing each chunk, the precision seems to remain stable. The ERDE percentages seem to present a similar pattern as for T1. After processing chunk 1, the highest amount of true positives are detected (48%), and again the highest amount of false positive cases are identified (56%).

Comparing our results with the ones of the models presented by other teams, based on the F1 Score, our best model got the seventh place among the 35 models presented. Taking into account the *ERDE writings* score, we can see that our model would have obtained the lowest value for the ERDE5 measure among all the lab results, with a percentage of 10.48%. Again, we assume that the other models were designed to see all the writings of the chunk from which they made a decision.

Table 3. Top ranked models regarding F1 score (T1 and T2).

Task	Model	F1	Precision	Recall	ERDE5 challenge	ERDE50 challenge	ERDE5 chunks	ERDE50 chunks	ERDE5 writings	ERDE50 writings
T1	UPFA	0.55	0.56	0.54	10.01%	8.28%	9.39%	7.35%	9.11%	6.41%
T2	UPFC	0.73	0.76	0.71	13.17%	11.60%	12.19%	9.74%	10.48%	8.17%

Table 4. Results obtained after processing each chunk (T1 and T2).

		Chunk									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
T1 - UPFA	**F1**	0.32	0.40	0.44	0.48	0.51	0.51	0.52	0.51	0.52	0.55
	P	0.43	0.49	0.52	0.54	0.55	0.55	0.55	0.53	0.54	0.56
	R	0.25	0.34	0.38	0.43	0.47	0.48	0.49	0.49	0.51	0.54
	ERDE5	9.26%	9.04%	9.04%	9.05%	9.06%	9.07%	9.08%	9.11%	9.11%	9.11%
	ERDE50	7.99%	7.16%	7.04%	6.72%	6.73%	6.62%	6.63%	6.65%	6.65%	6.41%
T2 - UPFC	**F1**	0.47	0.55	0.60	0.61	0.62	0.64	0.69	0.72	0.72	0.73
	P	0.74	0.75	0.77	0.75	0.73	0.74	0.76	0.76	0.76	0.76
	R	0.34	0.44	0.49	0.51	0.54	0.56	0.63	0.68	0.68	0.71
	ERDE5	10.92%	10.36%	10.36%	10.40%	10.44%	10.44%	10.44%	10.48%	10.48%	10.48%
	ERDE50	9.26%	8.68%	8.68%	8.72%	8.45%	8.45%	8.13%	8.17%	8.17%	8.17%

7 Conclusions and further work

In this paper we proposed several models for the early detection of cases of depression and anorexia, by dynamically processing users' text streams. Different machine learning approaches were designed using features extracted from the texts. These features were based on linguistic information, domain-specific vocabulary, and psychological processes. The models generated have a better performance for predicting anorexia. However, the results obtained have shown that the proposed approaches are suitable for the early detection of both depression and anorexia.

With the aim to improve our results, new features and learning algorithms will be tested. We plan to try other algorithms such as SVM, Neural Networks and voting methods, as they have been previously applied with promising results [15]. Features based on the posting time, topics and sentiment analysis are left to be tested. Finally, we will investigate how to avoid the prediction of too many false positives right after processing the first chunk.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

References

1. Agresti, A.: Categorical Data Analysis, pp. 165–167. Wiley Series in Probability and Statistics, Wiley (2013), https://books.google.es/books?id=UOrr47-2oisC
2. 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. 1 (2018). https://doi.org/10.1177/2167702617747074, https://doi.org/10.1177/2167702617747074

3. Arseniev-Koehler, A., Lee, H., McCormick, T., Moreno, M.: Proana: Pro-eating disorder socialization on twitter. Journal of Adolescent Health **58**, 659–664 (04 2016)
4. Bagroy, S., Kumaraguru, P., De Choudhury, M.: A social media based index of mental well-being in college campuses. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 1634–1646. CHI '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025909, http://doi.acm.org/10.1145/3025453.3025909
5. Benton, A., Mitchell, M., Hovy, D.: Multi-task learning for mental health using social media text. Proceedings of the 15th Conference of the EACL **abs/1712.03538**, 152–162 (2017), http://arxiv.org/abs/1712.03538
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (Oct 2001). https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324
7. Coppersmith, G., Harman, C., Dredze, M.: Measuring post traumatic stress disorder in twitter. In: Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014. pp. 579–582 (01 2014)
8. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in twitter. In: Association for Computational Linguistics Workshop of Computational Linguistics and Clinical Psychology. pp. 51–60. Baltimore, Maryland USA (2014)
9. 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. Denver, Colorado (2015)
10. De Choudhury, M., Counts, S., Horvitz, E.J., Hoff, A.: Characterizing and predicting postpartum depression from shared facebook data. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work; Social Computing. pp. 626–638. CSCW '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2531602.2531675, http://doi.acm.org/10.1145/2531602.2531675
11. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: ICWSM. pp. 1–10. AAAI (July 2013), https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/
12. Elberrichi, Z.: Text mining using n-grams. In: Proceedings of the International Conference on Computer Science and its Applications. pp. 25–38 (01 2006)
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). https://doi.org/https://doi.org/10.1016/j.cobeha.2017.07.005, http://www.sciencedirect.com/science/article/pii/S2352154617300384, big data in the behavioural sciences
14. Hwang, J.D., Hollingshead, K.: Crazy mad nutters: The language of mental health. In: Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 52–62. San Diego, California (2016)
15. Leiva, V., Freire, A.: Towards suicide prevention: Early detection of depression on social media. In: Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S., McMillan, D. (eds.) Internet Science. pp. 428–436. Springer International Publishing, Cham (2017)

16. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 28–39. Springer International Publishing, Cham (2016)
17. Losada, D.E., Crestani, F., Parapar, J.: CLEF 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In: Linda Cappellato, Nicola Ferro, L.G.T.M. (ed.) Conference and Labs of the Evaluation Forum. CEUR-WS.org (2017)
18. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
19. 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)2012. pp. 1–8. Beijing, China (01 2012)
20. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The Development and Psychometric Properties of LIWC2007. This article is published by LIWC Inc, Austin, Texas 78703 USA in conjunction with the LIWC2007 software program., http://www.liwc.net/LIWC2007LanguageManual.pdf
21. PreoŢiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H.A., Ungar, L.: The role of personality, age, and gender in tweeting about mental illness. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 21–30. Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/W15-1203, http://www.aclweb.org/anthology/W15-1203
22. Prieto, V.M., Matos, S., Álvarez, M., Cacheda, F., Oliveira, J.L.: Twitter: A good place to detect health conditions. PLOS ONE **9**(1), 1–11 (01 2014). https://doi.org/10.1371/journal.pone.0086191, https://doi.org/10.1371/journal.pone.0086191
23. Reece, A.G., Reagan, A.J., Lix, K.L.M., Dodds, P.S., Danforth, C.M., Langer, E.J.: Forecasting the onset and course of mental illness with twitter data. In: Nature: Scientific Reports. pp. 1–11 (2017)
24. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.H.: Towards assessing changes in degree of depression through facebook. In: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 118–125. Baltimore, Maryland USA (2014)
25. Seabrook, M.E., Kern, L.M., Rickard, S.N.: Social networking sites, depression, and anxiety: A systematic review. JMIR Ment Health **3**(4), 50 (Nov 2016). https://doi.org/10.2196/mental.5842, http://mental.jmir.org/2016/4/e50/
26. TJ, M., DR, M.: Adult utilization of psychiatric drugs and differences by sex, age, and race. JAMA Internal Medicine **177**(2), 274–275 (2017). https://doi.org/10.1001/jamainternmed.2016.7507, + http://dx.doi.org/10.1001/jamainternmed.2016.7507
27. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 3187–3196. CHI '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702280, http://doi.acm.org/10.1145/2702123.2702280

28. Wang, T., Brede, M., Ianni, A., Mentzakis, E.: Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 91–100. WSDM '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3018661.3018706, http://doi.acm.org/10.1145/3018661.3018706