Deep learning architectures and strategies for early detection of self-harm and depression level prediction

Ana-Sabina Uban[1,2] and Paolo Rosso[1]

[1] PRHLT Research Center, Universitat Politècnica de València
[2] Human Language Technologies Research Center, University of Bucharest
ana.uban+prof@gmail.com, prosso@dsic.upv.es

Abstract. This paper summarizes the contributions of the PRHLT-UPV team as a participant in the eRisk 2020 tasks on self-harm detection and prediction of depression levels from social media. Computational methods based on machine learning and natural language processing have a great potential to assist with early detection of mental disorders of social media users, based on their online activity. We use multi-dimensional representations of language, and compare various deep learning models' performance, exploring rarely approached avenues in previous research, including hierarchical deep learning architectures and pre-trained transformers and language models.

Keywords: deep learning · mental disorders · BERT · hierarchical attention network · self-harm · depression.

1 Introduction

Mental health disorders affect hundreds of millions of people worldwide; [17] depression alone is a major factor for suicide, and is usually underdiagnosed and undertreated. People affected by mental disorders often turn to social media to talk about their problems. There is an important opportunity for automatic processing of social media data in order to identify changes in mental health status that may otherwise go undetected before they develop more serious health consequences. Identifying people who start to develop signs of a mental illness early on is very important to managing its evolution, and in certain cases it can be life-saving. Recently, the recent COVID-19 pandemic is expected to exacerbate this problem, affecting mental health as well as physical health [9].

The CLEF eRisk Lab [3], organized every year since 2017, is dedicated specifically to identifying early signs of mental disorders from a user's social media posts, before the user was diagnosed with the disorder, for disorders including

[3] https://erisk.irlab.org/

depression, anorexia and thoughts of self-harm [10–12]. Each year a new task is organized around predicting a specific disorder: in 2017 and 2018 the shared tasks focused on depression detection, in 2019 a new task for anorexia prediction was organized, as well as a second task around predicting self-harm tendencies without any training data; in 2020 self-harm detection was again the topic, this time in a supervised setting. Datasets are collected from Reddit posts and comments selected from specific relevant sub-reddits, annotated by automatically detecting self-stated diagnoses of users. Healthy users are selected from participants in the same sub-reddits, thus making sure the gap between healthy and diagnosed users is not trivially detectable. For the self-harm task, the dataset includes only posts published before any involvement in the self-harm related communities, which conditions any model trained on this data to be capable of very early prediction, and at the same time adds difficulty to the task.

The language used by a speaker has been shown to contain strong indicators of an altered mental state. These can manifest both explicitly, at the level of the topics approached, or implicitly, at the level of the emotional charge of the text (greater negative emotion [5]), or even more subtle stylistic indicators (such as the increased use of first-person pronouns [25]). Textual data from social media, as a very rich and relatively easy to obtain type of data, as well as continuously growing source of real-time information, can thus be leveraged to gain many valuable insights into an individual's behavior and mental state and its evolution.

Most previous research related to automatic mental disorder detection from social media data have focused the study of depression [6, 8, 1], but other mental illnesses have also been studied, including generalized anxiety disorder [23], schizophrenia [13], post-traumatic stress disorder [3, 4], risks of suicide [16], anorexia [11] and self-harm [11]. The majority of studies on mental disorder detection use simple machine learning models (such as support vector machines (SVMs) and logistic regression) [6, 5]. Few studies have used more complex deep learning methods [21, 25, 26, 22]. At the level of features, most previous works have used traditional bag of words n-grams [3], as well as hand-crafted lexicons [24], LIWC features [5], or Latent Semantic Analysis [20, 24]. There are few studies which jointly consider several aspects of the language [22, 23].

This study summarizes our contributions as participants to the eRisk shared tasks on self-harm detection and assessment of depression levels [12]. We explore the use of deep learning for detecting mental disorders from text data, and compare various architectures, including hierarchical attention networks and transformers. We model our text data using a multi-aspect representation, through using features that reflect various complementary levels of the language, including content, style and emotion. For predicting the level of depression, we use traditional machine learning models including SVMs and Logistic Regression.

2 Task 1. Self-harm detection

The first task in eRisk 2020 consists of detecting whether a user is at risk of developing self-harm tendencies. Training data collected from Reddit was available, consisting of 340 users (of which 41 were labelled as positive) and their Reddit post history. Test data was provided as a stream of user posts, and candidate systems were asked to provide a decision (a binary number: a user is at risk or not), as well as a risk score (a real number), at each time step in the stream.

We participated in the task with five different models. We implement several neural network architectures, as well as experiment with pre-trained models and strategies for sampling training data in order to improve results. Details of the architectures used and the experimental setup are described below.

2.1 Features

Content features. We include a general representation of text content by transforming each text into word sequences. Preprocessing of texts includes lowercasing and tokenizing, removing punctuation and numbers; function words are not excluded. Most frequent 20,000 words were selected to form the vocabulary, and words not in the vocabulary were represented as a special "unknown" token. When passed as input to the neural networks, words within a sequence were encoded as embeddings of dimension 100. In order to initialize the weights of the embedding layers, we started from GloVe embeddings pre-trained on Twitter data. The choice of pre-trained embeddings was justified by their dimension, which is smaller than for other GloVe embeddings pre-trained on large corpora, leading to fewer mode parameters overall (in view of avoiding overfitting problems). Nevertheless, even though the data for this task is also sampled from social media, the two platforms (Reddit and Twitter) have significant differences as well; exploring the use of other embedding initializations (especially using embeddings pre-trained on longer texts) would be interesting in future experiments.

Style features. We aim at representing the stylistic level of texts through including function word and pronoun features. Function words have traditionally been used as stylistic markers, whereas increased use of pronouns, especially first person pronouns, has been shown to correlate with mental disorder risk [24]. We include two separate stylistic features: firstly, we extract from each text a numerical vector representing function words frequencies as bag-of-words. Separately, we include a simple scalar feature meant to capture the first person personal pronoun usage, by measuring the proportion of first person pronouns relative to the total number of words used in a text. We complement these with features extracted from the LIWC lexicon, as described below.

LIWC features. The LIWC[4] [18] is a lexicon mapping words in the English vocabulary to lexico-syntactic features of different kinds. It has been widely used in computational studies for analysing how suffering from mental disorders

[4] http://www.liwc.net/

manifests in an author's writings. LIWC categories have the capacity to capture different levels of language: including style (through syntactic categories), emotions (through affect categories) and topics (through content-oriented categories such as words referring to cognitive or analytical processes, or words referring to topics such as money, health or religion). We include in our analysis all 64 categories in the lexicon, and represent them as numerical vectors by computing for each category the ratio of words in a text that are related to the category, according to the lexicon.

Emotions and sentiment. We dedicate a few features to represent emotional content in our texts, since the emotional state of a user is known to be highly correlated with his/her mental health. Several of the LIWC categories aim to capture sentiment polarity and emotion content (*negative emotion*, *positive emotion*, *affect*, *sadness*, *anxiety*). We additionally include a second lexicon: the NRC emotion lexicon [14], which is dedicated exclusively to emotion representation, containing 9 different emotion categories: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *negative*,*positive*, *sadness*, *surprise*, *trust*. We represent NRC features similarly to LIWC features, by computing for each category the proportion of words in the text which are associated with that category.

2.2 Experimental setup

During the training phase as well as for testing, we do not consider social media posts individually as datapoints, since they are too short to be sufficiently predictive. Instead, we generate our datapoints by grouping sequences of 50 chronologically consecutive posts into larger chunks, to obtain more consistent samples of text as our datapoints. Features are computed at chunk-level.

As a consequence, prediction is always done on chunks of 50 posts. When analyzing the input stream of test data, we form new chunks of the last 50 posts received periodically (after every 20 new posts), and feed them to the networks to generate predictions.

As we will describe in the following section, we use two types of architectures for modelling the input: sequential and hierarchical. We adopt a special strategy for predictions on the first 50 posts in the stream: we pad the input data up to the size used during training (512 words in the case of the sequential setup and 50 posts of length 256 in the case of the hierarchical setup), but only submit the output score provided by the network, and as decisions (user is at risk or not) we submit zeros regardless of the output score, so as not to send premature alerts (since once a user is declared at risk, the decision can not be reverted).

Sequence sampling. For one of our runs, we employ a special strategy during the training phase. We attempt to augment the training data through generating "artificial" chunks of user posts, aside from the ones formed naturally through chunking the user's post history in chronological order. We do this by sampling from the post history randomly, following an exponential distribution so as to sample with higher probability from recent posts (which are more likely to contain signs of the disorder). The chronological order of posts is maintained.

Rolling average of predictions. As previously mentioned, for most runs predictions are generated using the last 50 posts seen in the test data stream. For one of our runs, we use a different strategy, by computing a rolling average of the most recent 3 network outputs: in this way, we hope to obtain more robust results that are not dependent only on the last batch of 50 user posts, but take into account a larger window of context.

2.3 Architectures

BiLSTM with attention. The first model we consider is a bidirectional LSTM network, with attention. Input word sequences are truncated at maximum 512 words, with words encoded as embeddings, and passed as input to the BiLSTM layer with 256 units, which is then fed to an attention layer. The bag-of-words features representing function word distribution are passed through a dense layer of 20 units; and the remaining extracted features (including pronoun, emotion and LIWC category usage) are concatenated into one vector. The output of the BiLSTM is concatenated along with the other features and the final representation passed through an output layer that generates the final prediction.

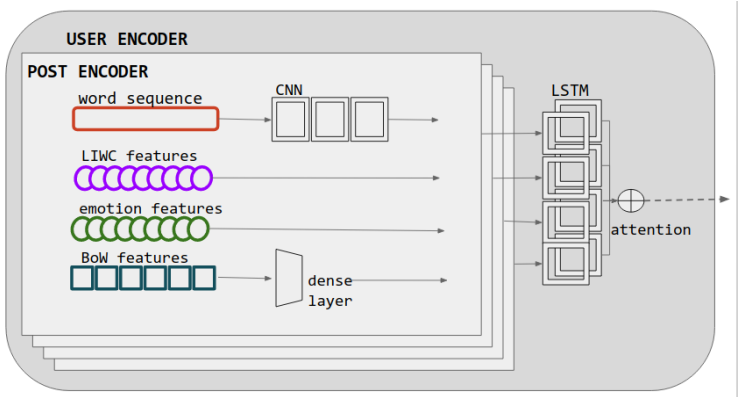

Fig. 1. Hierarchical attention network architecture.

Hierarchical Attention Network. Hierarchical attention networks (HAN) were introduced in [27] where they were used for review classification, by representing a text as a hierarchical structure where a document is comprised of sentences and a sentence is comprised of words. We propose that social media data in our setup is very well suited to such a hierarchical representation; in our case the hierarchy consists of user post histories, which are composed of social media posts, which are in turn composed of word sequences. Especially since the evolution of the mental state of a user is in itself a relevant indicator for the development of a disorder, as shown in [19], user-level representations are

expected to be natural and useful for modelling this problem. One other study has included post-level and user-level attention on their classifier's architecture, obtaining top results in the anorexia detection shared task [15].

In the hierarchical setup, posts within a chunk (datapoint) are stacked to form a hierarchical structure: word sequences (truncated at 256 words), as well as the rest of vectorial numerical and bag-of-words features, are stacked to form bi-dimensional vectors. Bag-of-words and numerical features also follow a hierarchical structure, with a set of features extracted for each post in the group, and stacked together into bi-dimensional vectors. The hierarchical network is composed of two components: a *post-level encoder*, which produces a representation of a post, and a *user-level encoder*, which generates a representation of a user's post history. For encoding the word sequence at post-level, we use a convolutional layer with 100 filters of length 3. Each of the posts in the input datapoint is encoded with the post-level encoder, and then they are stacked to form a bi-dimensional representation, which is then concatenated with the other features, and passed to the user-level encoder. We choose to model the user-level encoder as an LSTM layer with attention, with 32 units. The output of the user encoder is connected to the output layer which generates the final prediction. A depiction of the hierarchical architecture is shown in Figure 1.

Transformers. We experiment with state-of-the-art language models based on transformer architectures, which have been shown to obtain high performances on a wide range of NLP tasks, with minimal task-specific training. We use pre-trained BERT [7] models for English (the "base" versions of the models) with one trainable output layer and fine-tune them for our task.

Ensemble. Finally, we use a simple ensemble model for one of our runs: predictions are generated through averaging the outputs of several other models on the received input.

2.4 Models submitted

The models and setups used for each of the five runs submitted by our team are described below.

Run 0. BERT + sequence sampling. For our first run we used the pre-trained and fine-tuned BERT model. During fine-tuning, the sequence sampling strategy for data augmentation was used.

Run 1. BiLSTM. Run 1 consists of the BiLSTM model described in the previous section.

Run 2. Hierarchical CNN + LSTM. For run 2 we used the hierarchical attention network with CNN and LSTM layers. Due to memory limitations, we only generated predictions for the first 50 posts in the stream: all subsequent predictions (for all datapoints in the stream) were based on these outputs.

Run 3. Ensemble. For this run, we used an ensemble of the first three models: BERT, the BiLSTM and the hierarchical attention network. To obtain prediction scores, we averaged the outputs of the three networks for each input datapoint. A user is considered at risk if the obtained output exceeds the 0.5 threshold.

Run 4. Rolling average of BiLSTM. For our last run, we used the rolling average strategy described in the previous section, to obtain a smoothed version of the model's outputs. For each timestep, we averaged the output of the BiLSTM model for the most recent three inputs (chunks of 50 posts).

2.5 Results

Table 1 show the official results obtained for each of our runs. Evaluation measures included the traditional precision, recall and F1-scores computed at user-level, as well as some metrics specifically designed for measuring how early risk was detected: latency-weighted F1, which is the F1-score weighted by a penalizing factor for late predictions, and *ERDE* [12], a measure of error that increases when predictions are delayed. For comparison, we include the systems that obtained best scores for each metric.

Run	Precision	Recall	F1	$ERDE_5$	$ERDE_{50}$	latencyw-F1
BERT+seq-sampling	.469	.654	**.546**	.291	.154	.462
BiLSTM	.710	.212	.326	.251	.235	.172
HAN	.271	.577	.369	.339	.269	.298
Ensemble	.846	.212	.338	.248	.232	.178
BiLSTM+rolling	.765	.375	.503	.253	.194	.423
iLab/run 1	**.913**	.404	.560	.248	.149	.540
SSN_NLP/run 1	.283	**1**	.442	.205	.158	.442
iLab/run 4	.828	.692	**.754**	.255	.255	.476
iLab/run 2	.544	.654	.594	**.134**	.118	.592
iLab/run 3	.564	.885	.689	.287	**.071**	.572
iLab/run 0	.833	.577	.682	.252	.111	**.658**

Table 1. Official results for task 1

Run	P@10	NDCG@10	NDCG@100
BERT+seq-sampling	**1**	**1**	.68
BiLSTM	.9	.81	.75
HAN	.6	.69	.48
Ensemble	.9	.81	.75
BiLSTM+rolling	.9	.90	.69
iLab/run 3	**1**	**1**	**.84**

Table 2. Ranking metrics for task 1

The best F1-scores were obtained with the BERT model using sequence sampling training, showing that pre-trained transformers are powerful for external

tasks including detection of self-harm, and also that the sequence sampling strategy might be an effective method for data augmentation. The second best results were obtained with the last model - the rolling average of outputs strategy brings significant improvement to predictions compared to the base model (simple BiLSTM). We attribute the poorer performance of the HAN and ensemble models to the small size of test data used for predictions (first 50 posts in the stream).

A second evaluation approach treats the task as a ranking task, by using the system's continuous risk scores and ranking users in order of risk according to these scores. Metrics specific to ranking tasks are used to measure performance, including precision @ k (P@10), and Normalized Discounted Cumulative Gain @ k (NDCG@10, NDCG@100). In Table 2 we show the evaluation results for our systems using the ranking metrics, measured on the first 500 posts in the input stream. Our models perform well on these metrics, the first system obtaining perfect scores for both metrics measured @ 10. For comparison, we include the system that obtained best scores in terms of all ranking metrics @ 500 writings, submitted by the iLab team.

3 Task 2. Predicting levels of depression

The second task consisted of predicting the level of depression of social media users, by predicting answers to a 21-question questionnaire for assessing depression, where each question can have one of four to six answers. Training data consisting of 20 labelled users was available beforehand. The test data consisted of 70 users' social media posts, and the participating systems had to predict their answers to each of the questions.

Several evaluation metrics were used, measuring how well the predictions match the true labels, from more fine-grained to more general levels, including: average hit rate (AHR), average closeness rate (ACR), average difference between overall depression levels (ADODL), depression category hit rate (DCHR).

We participated with three different models in this task. The details of the models and features used are described below.

3.1 Features

For the first two models, we used a few of the same features described in the previous sections. We included the lower-dimensionality numerical features: LIWC and emotion categories, represented as continuous vectors. For obtaining user-level representations, we averaged the values of these vectors computed for each of the user's posts. Since it has been shown that the evolution of certain behaviors and linguistic markers is in itself predictive of developing a disorder or not, we choose to capture the variation of the features extracted, by including in our feature vectors the standard variations (aside from the averages) seen in the distribution of each feature across a user's history of posts.

For our final model, we tried to leverage pre-trained language models in order to obtain semantic representations of the user's social media posts. To this effect,

we extracted sentence representations from Universal Sentence Encoder (USE) [2] for each of the posts in a user's history, obtaining a continuous vectorial representation for each post. A user's representation was obtained by averaging the representations of each of his/her posts.

3.2 Models

We chose to use simpler traditional machine learning models for this task with fewer parameters than the neural networks used in task 1, to suit the small size of the training data: we experimented with SVM and logistic regression models, using the features previously described.

All models were trained on the available training data, and the trained models were used to make predictions on the new data in the testing phase. For each of the models, we modelled the task as a multi-label multi-class classification problem, by training one model for each of the 21 questions, where each question can be assigned one of 4-6 labels (depending on the question).

LogReg-features The first model used was a logistic regression model with the lexicon-based features represented as numerical vectors.

SVM-features For the second run, we used an SVM with RBF kernel, with the same features as for the previous run.

SVM-USE Our last model was an SVM with RBF kernel, and USE features.

3.3 Results

Run	AHR	ACR	ADODL	DCHR
LogReg-features	34.01%	67.07%	80.05%	**35.71%**
SVM-features	34.56%	67.44%	80.63%	**35.71%**
SVM-USE	36.94%	69.02%	81.72%	31.53%
BioInfo@UAVR	**38.30%**	69.21%	76.01%	30.00%
iLab run2	37.07%	**69.41%**	81.70%	27.14%
relai_lda_user	36.39%	68.32%	**83.15%**	34.29%

Table 3. Official results task 2

Table 3 shows official results results for task 2, for all evaluation metrics. Our best models in terms of DCHR were the models using lexicon-based features, which obtained the maximum score of all participating teams on this metric. The model using USE features has better performance than the other two for the rest of the metrics. The good scores obtained with simple models and features suggest the problem may not be well suited to complex representations and architectures, possibly due to the small size of the training data. For comparison, we include in the table results of the systems that obtained best scores in terms of the other metrics (aside from DCHR). Overall, scores for this task were modest for all

participating teams, suggesting predicting the level of depression is a difficult task.

4 Conclusion

In this paper we presented the contributions of the PRHLT-UPV team in the eRisk 2020 shared tasks: self-harm detection and the prediction of depression levels, based on social media text data. We used multi-dimensional features to represent various levels of the language, including content, style and emotion. In the first task, where more training data was available, we experimented with different deep learning architectures, including hierarchical attention networks and transformers, as well as with different strategies concerning the experimental setup: such as sequence sampling for data augmentation, and rolling average for smoothing model outputs. For the second task we used traditional models such as SVM and logistical regression, with features including style and emotion features, as well as semantic sentence representations from pre-trained language models. We obtained best scores in terms of detecting the general depression category in the second task.

Acknowledgements

The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

References

1. Abd Yusof, N.F., Lin, C., Guerin, F.: Analysing the causes of depressed mood from depression vulnerable individuals. In: Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017). pp. 9–17 (2017)
2. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
3. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. pp. 51–60 (2014)
4. 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)
5. 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 (2014)

6. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Seventh international AAAI conference on weblogs and social media (2013)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D.A., Schwartz, H.A.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences **115**(44), 11203–11208 (2018)
9. Lee, S.A., Mathis, A.A., Jobe, M.C., Pappalardo, E.A.: Clinically significant fear and anxiety of covid-19: A psychometric examination of the coronavirus anxiety scale. Psychiatry Research p. 113112 (2020)
10. 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)
11. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2019 early risk prediction on the internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 340–357. Springer (2019)
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: A. Arampatzis, E. Kanoulas, T.T.S.V.H.J.C.L.C.E.A.N.L.C.N.F.e. (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). Springer International Publishing (2020)
13. Mitchell, M., Hollingshead, K., Coppersmith, G.: Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality. pp. 11–20 (2015)
14. Mohammad, S.M., Turney, P.D.: Nrc emotion lexicon. National Research Council, Canada **2** (2013)
15. Mohammadi, E., Amini, H., Kosseim, L.: Quick and (maybe not so) easy detection of anorexia in social media posts. In: CLEF (Working Notes) (2019)
16. O'dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., Christensen, H.: Detecting suicidality on twitter. Internet Interventions **2**(2), 183–188 (2015)
17. Organization, W.H.: Depression: A global crisis. world mental health day, october 10 2012. World Federation for Mental Health, Occoquan, Va, USA (2012)
18. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71**(2001), 2001 (2001)
19. Ragheb, W., Azé, J., Bringay, S., Servajean, M.: Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In: CLEF (Working Notes) (2019)
20. Resnik, P., Garron, A., Resnik, R.: Using topic modeling to improve prediction of neuroticism and depression in college students. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1348–1353 (2013)
21. Sadeque, F., Xu, D., Bethard, S.: Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection. In: CEUR workshop proceedings. vol. 1866. NIH Public Access (2017)

22. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.S., Zhu, W.: Depression detection via harvesting social media: A multimodal dictionary learning solution. In: IJCAI. pp. 3838–3844 (2017)
23. Shen, J.H., Rudzicz, F.: Detecting anxiety through reddit. In: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality. pp. 58–65 (2017)
24. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In: CLEF (Working Notes) (2017)
25. 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)
26. Wang, Y.T., Huang, H.H., Chen, H.H.: A neural network approach to early risk detection of depression and anorexia on social media text. In: CLEF (Working Notes) (2018)
27. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 1480–1489 (2016)