Quick and (maybe not so) Easy Detection of Anorexia in Social Media Posts

Elham Mohammadi, Hessam Amini, and Leila Kosseim

Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, QC H3G 2W1, Canada
{elham.mohammadi,hessam.amini,leila.kosseim}@concordia.ca

Abstract. This paper presents an ensemble approach for the early detection of anorexia in social media posts. The approach utilizes several attention-based neural sub-models to extract features and predict class probabilities, which are later used as input features to a Support Vector Machine (SVM) making the final classification. The model was evaluated on the first task of eRisk 2019, whose aim was the early detection of anorexia in Reddit posts. Our submission, named *CLaC* achieved F1 and latency-weighted F1 scores of 0.7073 and 0.6908 respectively, allowing it to rank first in terms of these metrics, and achieved competitive results based on other evaluation metrics.

Keywords: Anorexia · Early detection · Social media · Ensemble classifier · Neural networks · Support vector machine

1 Introduction

In the last decade, the use of social media to express personal thoughts, emotions, and ideas has become more and more prevalent. The analysis of online data can be useful for many purposes, such as business and marketing, political planning, prediction of stock market [10], as well as enhancing awareness of emergencies [34]. Another noteworthy line of research has focused on the detection of toxicity, hate speech, aggression and cyber bullying on online platforms, an effort that could facilitate timely interventions in violent situations [8,31].

In healthcare applications, online posts have been used for detecting disease outbreaks [23], finding smoking patterns [30], and the identification of adverse drug reactions [33]. Another useful application is the automatic detection of mental health issues, a relatively recent field which has attracted the attention of many researchers in Natural Language Processing (NLP). Corpora from Twitter, Facebook, blogs and online forums, and Reddit are used as resources to detect various mental health problems, such as anxiety, depression, suicide ideation, and eating disorders [3].

Although automatically monitoring online forums to detect cases of mental health issues is beneficial, the elapsed time between the first signs of a mental issue and the actual detection of a potential victim can play a crucial role. Earlier detection of a harmful behavior can help moderators better handle the situation. However, to the best of our knowledge, not much research has specifically addressed the task of early detection of mental health issues.

The eRisk shared task [17] was created with the goal of addressing issues related to early risk detection of mental health problems. According to [17], early detection can be useful in many applications from the identification of potential sexual offenders to the detection of victims of suicidal tendencies, making intervention possible before it is too late. [17] argues that while current risk assessment approaches often aim at detecting harmful behavior after the fact, it is very important to consider the timing of risk detection and to minimize the time between the observation of the first evidence of destructive behavior and triggering an alarm. To that end, the organizers of the eRisk shared task have encouraged the development of approaches which model the process rather than the outcome, as well as developing reliable evaluation metrics and test collections tailored to early risk detection.

The aim of this work is to propose of a model for the early detection of anorexia and to evaluate it using the eRisk 2019 data and evaluation metrics [19].

The rest of the paper is organized as follows: Section 2 provides an overview of the related literature. Section 3 consists of a brief summary of the task and the data set used. Section 4 presents the general model architecture that has been developed. Section 5 is dedicated to a more detailed description of model variants that were employed for the experiments. Section 6 includes a summary and discussion of the results. Section 7 concludes the paper and presents some interesting future directions.

2 Related Work

Many researchers have used corpora from Twitter, Facebook, Reddit, blogs and online forums as resources to experiment with classification tasks pertaining to mental health issues [3].

Pestian et al. [27] experimented with different machine learning methods for suicide note classification. The features used in the study included words, part of speech tags, readability scores, and emotions. The best accuracy of 74% was achieved by a logistic regression model.

DeVault et al. [9] studied the symptoms of psychological distress in dialogues with a virtual agent. The use of a Naïve Bayes classifier for the detection of post-traumatic stress disorder (PTSD) and distress yielded a 20% improvement over the baseline accuracy of 53.5%, and showed that the automatic assessment of psychological distress is indeed possible.

More recently, Jackson et al. [13] used clinical texts obtained through the Clinical Record Interactive Search[1] to extract symptoms of severe mental illness. The authors made use of TextHunter [1] (a natural language processing information extraction tool) and an SVM classifier, and were able to classify 38 symptoms with an F1-score of 85%.

Shen and Rudzicz [28] used different feature sets including word2vec embedding, latent Dirichlet allocation topic modelling, lexico-syntactic features, and n-grams (unigrams and bigrams) to detect anxiety in Reddit posts. Initially, the authors compared the results achieved by an SVM and a 2-layer neural network. Though both classifiers performed well, the SVM yielded marginally better results. However, they achieved their best result of 98% accuracy using the neural network with n-gram probabilities and word embeddings combined with Linguistic Inquiry and Word Count (LIWC) features.

Coppersmith et al. [6] explored the automatic detection of post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD) in Twitter data, using LIWC features and character and word n-grams, and found the latter resulting in superior performance.

Benton et al. [2] used multi-task learning to predict suicide risk and a variety of mental health conditions from Twitter data, including anxiety, depression, PTSD, and schizophrenia. It was found that a multi-task framework can be effectively used in cases with limited data.

Apart from individual efforts, shared tasks (e.g. [7,21,20,35]) have also been organized to encourage the development of common benchmarks (datasets and metrics) and the comparison of approaches for the detection of distress in online textual data.

All of the previous work described above used a classic classification approach that does not measure how early the detection is performed. On the other hand, the eRisk shared tasks [17,18,19] focus on the early detection of mental health issues. In the first edition of eRisk [17], the data set used was a collection of social media posts and comments from depressed and non-depressed authors, recorded chronologically. As evaluation metric, Early Risk Detection Error (ERDE) was used, an error measure which assigns a penalty to late decisions and rewards early ones [16]. As it was the first edition of this shared task, many teams focused on making accurate rather than early decisions, with the highest F1-score being 64% and the lowest $ERDE_{50}$ score[2] being 9.68% [17].

The second eRisk shared task [18] included two tasks: Early risk detection of depression and early risk detection of anorexia. Like the year before, the ERDE evaluation metric was used as the main metric alongside F1, precision, and recall [18]. The best performing systems, in both tasks, were designed by Trotzek et al. [32]. Their team experimented with different variations of bag of words features and a Convolutional Neural Network (CNN) [15] as well as ensemble models. In the depression task, their system achieved an F1-score of 64% and an $ERDE_{50}$ of

[1] `https://crisnetwork.co`

[2] A detailed description of $ERDE_o$, where o is either 5 or 50, can be found in [18].

6.44%. In the anorexia task, they achieved an F1-score of 85% and an $ERDE_{50}$ of 5.96%.

In this work, we present an ensemble approach that can be used for the detection of different types of distress in textual data. We investigate the effectiveness of the model by presenting and analyzing our results in the first task of eRisk 2019 [19].

3 Task and Dataset

Following the success of the eRisk 2018 task 2 [18], the eRisk 2019 task 1 [19] focuses on the early detection of anorexia in online posts. The data used for the task is a collection of Reddit users labelled as anorexic or non-anorexic [16], along with a collection of their Reddit posts, recorded chronologically.

For the training phase, the data from the previous year (eRisk 2018 task 2), including both training and test sets, was made available. For the testing phase, posts were released on an item-by-item basis in chronological order for a new collection of Reddit users. The goal was to detect users suffering from anorexia, having observed as few posts from them as possible. As a result, in addition to precision, recall, and F1-score, two other metrics were used: Early Detection Error (ERDE) measure which penalizes late decisions, and latency-weighted F1, a modified version of F1 score that takes into account the delay of the decision[3].

Table 1 shows some statistics of the datasets. As shown in the table, the datasets are highly imbalanced, with about 90% of the users not suffering from anorexia.

Table 1. Distribution of user labels in the datasets. The 2018 datasets refer to the eRisk 2018 task 2 data.

Dataset	Source	Positive	Negative	All
Training	Train 2018	20 (13%)	132 (87%)	152
Validation	Test 2018	41 (13%)	279 (87%)	320
Testing	–	73 (9%)	742 (91%)	815

4 System Overview

Fig. 1 shows the architecture of the model that we used for the eRisk 2019 shared task. The full model includes 8 different neural sub-models, followed by a fusion component, which concatenates the neural features and predicted class probabilities from different sub-models, and forwards them to a final SVM classifier.

This section will provide a more detailed explanation of the different components of the model.

[3] The details of the evaluation metrics for eRisk 2019 task 1 is explained in [19].

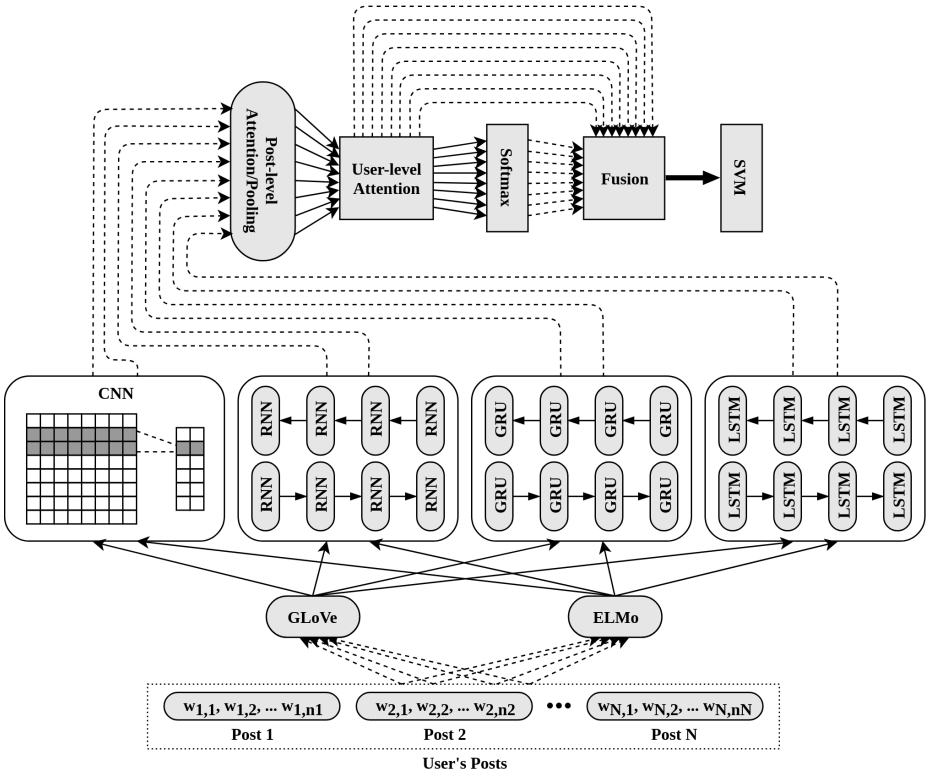

Fig. 1. Architecture of the model. The number of arrows between components corresponds to the number of sub-models that move in that flow. The rounded-corner boxes represent the components that work at the post level, while the sharp-corner boxes are user-level components. The solid lines represent neural connections; while the dotted lines show the flow of data without the existence of a neural connection. The bold arrow between the Fusion and SVM corresponds to the flow of data that exists only in the final model.

4.1 Sub-models

As shown in Fig. 1, each sub-model includes an input layer that receives as input the posts by a user, and vectorizes its tokens using an embedding layer. The output of the input layer is then fed to a hidden layer, which is followed by a post-level attention/pooling layer that creates a representation of the post from its constituent tokens. The user-level attention layer is responsible to calculate the vector representation of the user, using her/his online posts. Finally, the output (classification) layer predicts the probability distribution of the positive and negative classes (i.e. anorexic versus non-anorexic).

Our main focus during the development of the sub-models was to include diversity of information sources, so that the final ensemble model can incorporate different points of views when performing the final classification.

Input Layer. The inputs to the model are the online posts of each user. Each post is first tokenized, and the tokens are sent to the word embedder, in order to be converted into dense vectors. As shown in Figure 1, these token vectors are then fed to the hidden layer.

Two different pretrained word embeddings were experimented with. The first word embedder was the 300d version of GloVe [26] that was pretrained on 840B tokens of web data from Common Crawl. The second word embedder was the original 1024d version of ELMo, which was pretrained on the 1 Billion Word Language Model Benchmark [4]. These two word embeddings were used in order to provide our ensemble model with sub-models that utilize both contextual (ELMo) and non-contextual (GloVe) word embedders in their input layer.

Hidden Layer. The hidden layer is responsible for processing the token vectors, generated by the input layer. As shown in Fig. 1, we have experimented with four hidden architectures in our sub-models: a CNN [15] that processes token n-grams separately, and a Bidirectional Vanilla RNN (BiRNN), a Bidirectional Long Short-Term Memory (BiLSTM) [12] and a Bidirectional Gated Recurrent Unit (BiGRU) [5], all of which process token vectors sequentially, from first to last and vice-versa, by taking into account the preceding and following tokens, respectively.

Post-level Attention/Pooling Layer. Following [32], for the sub-models that use CNN in the hidden layer, a max pooling is applied to the outputs of the hidden layer after being passed through a Concatenated Rectified Linear Unit (CReLU, i.e. ReLU applied on the concatenation of each vector and its negative).

In the models that use BiRNN, BiLSTM, or BiGRU in their hidden layer, an attention mechanism is responsible for computing the representation of a post (P) by weighted-averaging over the outputs of the hidden layer for each token in the post, where the weights assigned to each token is calculated automatically. The function used by the attention mechanism can be shown in Equation 1:

$$P = \sum_{t=1}^{n} y_t \omega_t \tag{1}$$

where y_t represents the output of the recurrent hidden layer at time-step t, and ω_t is the weight assigned to the output in that time-step.

In our model, the attention mechanism uses an N-to-1 feed-forward layer (with the weights w, where N is equal to the size of the output vectors of the recurrent hidden layer) to map the output of the hidden layer at each time-step (e.g. y_t) to a scalar (e.g. ν_t):

$$\nu_t = y_t \times w \tag{2}$$

These scalars are then concatenated, and softmax is applied to the resulting vector. The resulting vector from the softmax will include the weights that are used by the attention mechanism:

$$\omega = Softmax([\nu_1, \nu_2, \nu_3, \ldots, \nu_n]) \tag{3}$$

User-level Attention Mechanism. Knowing that the posts by a user do not contribute equally to detect her/his mental state [32], a user-level attention mechanism is used to make the system learn to automatically detect the contribution of each post to the final classification of the user.

The mechanism of the user-level attention is similar to the post-level attention mechanism, but computes a vector representation of a user from the representation of her/his posts (resulted from the post-level attention/pooling).

Output (Classification) Layer. The final layer in the sub-models is a feed-forward fully-connected layer that maps the output of the user-level attention to a vector with size 2 (corresponding to the *negative* and *positive* classes). At the end of this layer, a softmax activation function gives as the output, the predicted probability distribution over the classes *negative* and *positive*.

4.2 Ensemble Model

As shown in Fig. 1, the ensemble model is composed of several neural sub-models, a fusion component, and a final SVM classifier. The fusion component concatenates the outputs of the user-level attention units (which will subsequently be referred to as neural features), and the predicted probability distributions of the two classes, resulting from the softmax activation functions from all its constituent sub-models. The output of the fusion component is taken as the final representation of a user. This representation is finally fed to an SVM classifier to perform the ensemble classification.

5 Experimental Setup

This section describes our experiments with the above model for our participation to the eRisk 2019 shared task [19].

5.1 Sub-models Implementation

PyTorch [24] was used to implement and train the sub-models. The Adam optimizer [14] was used, and the learning rate was set to 5×10^{-4}. Cross-entropy was used as the loss function, in order to handle the imbalanced distribution of the positive and negative classes in the training set (see Table 1), weights proportional to the inverse of the number of samples of each class were assigned to that class. Due to lack of computational resources, mini-batches with a maximum size of 128 were used at the post level for each user and only the first

100 tokens of the posts were used[4]. In order to minimize the amount of padding in the batches, posts with similar number of tokens were assigned to the same batch.

In order to fine-tune the other hyperparameters of the sub-models (including the number and size of convolutional filters, number of recurrent units, and number of training epochs), each sub-model was individually trained with training set and optimized on the validation set (see Table 1), based on F1 score. The specifics of the 8 different sub-models are shown in Table 2. Since each sub-model is composed of a unique pair of hidden layer and word embedding type, they will later be referred to as <*hidden-type*>-<*embedding-type*> (see the second column of Table 2).

Table 2. Hyperparameter values used in the 8 sub-models

#	Name	Hyperparameters
1	*CNN-GloVe*	100 bigram convolution filters, trained for 10 epochs
2	*CNN-ELMo*	200 unigram filters and 50 bigram convolution filters, trained for 6 epochs
3	*BiRNN-GloVe*	one layer of 64 vanilla RNN units, trained for 14 epochs
4	*BiRNN-ELMo*	one layer of 50 vanilla RNN units, trained for 13 epochs
5	*BiLSTM-GloVe*	one layer of 32 bidirectional LSTM units, trained for 31 epochs
6	*BiLSTM-ELMo*	one layer of 64 bidirectional LSTM units, trained for 14 epochs
7	*BiGRU-GloVe*	one layer of 64 bidirectional GRUs, trained for 14 epochs
8	*BiGRU-ELMo*	one layer of 64 bidirectional GRUs, trained for 8 epochs

5.2 Ensemble Classifiers

Scikit-learn [25] was used to develop the SVM classifier used in the ensemble model. Three different versions of ensemble classifiers were developed:

1. ***Ens-Feat*** is the version of the ensemble model that only utilizes the neural features. The SVM classifier in this version uses a sigmoid kernel. The γ and C parameters in the SVM were set to *auto* (i.e. 1/<number-of-features>) and 4, respectively.
2. ***Ens-Prob*** uses only the predicted class probabilities from the softmax activation function at the end of the neural sub-models. It utilizes a polynomial kernel with the degree of 1. The γ and C parameters in the SVM were set to *scale* (i.e. 1/[<number-of-features>×<variance-of-features>]) and 1, respectively.
3. ***Ens-All*** utilizes both neural features and predicted class probabilities in its SVM classifier, that uses a sigmoid kernel, and has its values of γ and C set to *auto* and 2, respectively.

[4] This limit only truncated a small number of posts, as the average length was ∼37.47 tokens in the eRisk 2018 task 2 data.

5.3 Submitted Runs

Based on the results with the validation set, 5 runs were submitted to the shared task server. For the 1st and 2nd runs, *CNN-GloVe* and *CNN-ELMo* were used, respectively, as stand-alone models[5], and *Ens-Feat*, *Ens-Prob*, and *Ens-All* comprised the 3rd, 4th and 5th runs.

6 Results and Discussion

Table 3 shows the official results of our submissions, as well as selected runs from other teams (as reported in [19]) that achieved the best result with one of the official evaluation metrics, or achieved competitive results. For the results of our team (*CLaC*), we indicate in Table 3 the specific name of the models used in the five submitted runs.

Table 3. Official results on the first task of the eRisk 2019 shared task. *#writings*: maximum number of writings (Reddit posts) that were processed for a user, *P*: Precision, *R*: Recall, *l-w F1*: Latency-Weighted F1 score.

team	model	run	#writings	P	R	F1	$ERDE_5$	$ERDE_{50}$	l-w F1
CLaC	CNN-GloVe	0	109	0.4463	0.7400	0.5567	0.0672	0.0393	0.5437
CLaC	CNN-ELMo	1	109	0.6061	0.8219	0.6977	0.0573	0.0312	0.6895
CLaC	Ens-Feat	2	109	0.6020	0.8082	0.6900	0.0602	0.0313	0.6766
CLaC	Ens-Prob	3	109	0.6292	0.7671	0.6914	0.0627	0.0355	0.6752
CLaC	Ens-All	4	109	0.6374	0.7945	**0.7073**	0.0625	0.0343	**0.6908**
lirmm		0	2024	0.74	0.63	0.68	0.09	0.05	0.63
lirmm		1	2024	**0.77**	0.60	0.68	0.09	0.06	0.62
Fazl		2	2001	0.09	**1.00**	0.16	0.17	0.11	0.14
UNSL		0	2000	0.42	0.78	0.55	**0.06**	0.04	0.55
UNSL		4	2000	0.31	0.92	0.47	0.06	**0.03**	0.46
INAOE-CIMAT		3	2000	0.67	0.68	0.68	0.09	0.05	0.63

As shown in Table 3, the model *Ens-All* achieved the highest F1 (0.7073) and latency-weighted F1 (0.6908) scores of all participants' runs. This was in line with our intuition that using an ensemble model that makes use of both neural features and predicted class probabilities from the 8 sub-models has a higher capability of detecting the correct class after observing a small number of writings. The results also show that the *CNN-ELMo* model can achieve F1 and latency-weighted F1 scores that are competitive to *Ens-All*, and outperforms *Ens-Feat* and *Ens-Prob* in these two metrics. The *CNN-ELMo* model also resulted in the best recall, $ERDE_5$ and $ERDE_{50}$, showing the potential of this model to be used independently for the task of early risk detection of anorexia.

Table 3, also shows that all our models, except *CNN-GloVe* (run 0) achieved significantly superior performances in terms of F1 score and latency-weighted

[5] These two sub-models achieved the most promising results among all the sub-models, during the training phase.

F1 (teams *lirmm* and *INAOE-CIMAT* achieved the next best F1 and latency-weighted F1 scores). Run 1 of team *lirmm* achieved the highest precision. The best recall was achieved by run 2 of the team *Fazl*. Runs 0 and 4 of the team *UNSL* achieved the highest $ERDE_5$ and $ERDE_{50}$, respectively, where we could achieve competitive results using *CNN-ELMo*.

The number of writings processed by the models submitted by each team shows that our models used a significantly lower number of writings in comparison to the other teams[6]. This shows that our systems have a great potential of making early and correct decisions. This is supported by an even larger gap between the latency-weighted F1 scores of our team and the runs submitted by other teams, in comparison to the gap in F1 score.

Although our systems achieved the best or competitive results according to different evaluation metrics, we suffered from lack of computational resources when running the models that use the ELMo embedder for around 2000 iterations. The models had to be run for approximately 2000 times due to the item-by-item release of the test data which was chosen for the eRisk 2019 shared task (in the previous eRisk shared tasks, the test data was released in 10 chunks, making the number of iterations equal to 10). Despite this technical drawback, the advantages of using ELMo to extract context-sensitive embeddings greatly outweigh its disadvantages. This can also be observed by comparing the results achieved by *CNN-GloVe* and *CNN-ELMo*.

7 Conclusion and Future Work

This paper presents an ensemble approach which can be used to detect distress in the social media posts of a user. The ensemble model utilizes neural features alongside predicted class probabilities which are output by 8 different neural sub-models. Using this model and under the team name *CLaC*, we participated to the first task of eRisk 2019 [19], which was aimed at the early detection of anorexia in online posts, and ranked first in terms of F1 and latency-weighted F1 scores.

Using a similar architecture, we also participated to the CLPsych 2019 shared task [22], whose aim was to assess suicide risk based on online posts. Considering that our ensemble model ranked first in tasks A and C of this shared task, the same model architecture seems applicable to other similar tasks, where the goal is to detect different types of mental health issues using social media posts.

We believe that the user-level attention mechanism has played an important role in the good results achieved on these shared tasks. It would be interesting to qualitatively analyze the results of the attention mechanism, to see how they correlate with human perception, i.e. whether the posts to which the attention mechanism assigns more weights are actually the same posts that seem more informative to a health specialist for detecting anorexia.

Also, during the development phase, it was found that removing each of the 8 sub-models (evens the sub-models with low individual performances) negatively

[6] The average number of writings processed by the participating teams was 1273.

affected the result of the final ensemble classifier. It would be interesting to measure quantitatively the contribution of each of the 8 neural sub-models in the result of the final classifier. This could then be leveraged to improve the performance of the system.

An additional research direction is the use of linguistic features and metadata. The current model does not explicitly use such features, however Trotzek et al. [32] showed that they can significantly improve early detection of anorexia.

Lastly, it would be interesting to experiment with more diverse architectures in the neural sub-models (e.g. by using other hidden layer architectures, such as recursive neural networks [11,29]) as a way of improving the performance of the current ensemble classifier.

Acknowledgment

We would like to thank the reviewers for their comments on an earlier version of this paper.

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Ball, M., Patel, R., Hayes, R.D., Dobson, R.J., Stewart, R.: TextHunter – a user friendly tool for extracting generic concepts from free text in clinical research. In: AMIA Annual Symposium Proceedings. vol. 2014, p. 729. American Medical Informatics Association (2014)
2. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017). pp. 152–162. Association for Computational Linguistics, Valencia, Spain (April 2017)
3. Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering **23**(5), 649–685 (2017)
4. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T.: One billion word benchmark for measuring progress in statistical language modeling. In: 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014). Singapore (September 2014)
5. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). pp. 1724–1734. Doha, Qatar (October 2014)
6. 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 (CLPsych 2014). pp. 51–60. Baltimore, Maryland, USA (June 2014)

7. 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 (CLPsych 2015). pp. 31–39. Association for Computational Linguistics, Denver, Colorado (2015)
8. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the Eleventh International Conference on Web and Social Media. pp. 512–515. Montréal, Canada (May 2017)
9. DeVault, D., Georgila, K., Artstein, R., Morbini, F., Traum, D., Scherer, S., Morency, L.P., et al.: Verbal indicators of psychological distress in interactive dialogue with a virtual human. In: Proceedings of the Special Interest Group on Discourse and Dialogue Conference (SIGDIAL 2013). pp. 193–202 (2013)
10. Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C.: Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications **69**, 214–224 (2017)
11. Goller, C., Kuchler, A.: Learning task-dependent distributed representations by backpropagation through structure. Neural Networks **1**, 347–352 (1996)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
13. Jackson, R.G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R.J., Stewart, R.: Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (CRIS-CODE) project. British Medical Journal (BMJ open) **7**(1) (2017)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: The 3rd International Conference for Learning Representations (ICLR 2015). San Diego, California, USA (May 2015)
15. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision, pp. 319–345 (1999)
16. 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. Evora, Portugal (September 2016)
17. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In: Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 346–360. Dublin, Ireland (September 2017)
18. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: Early Risk Prediction on the Internet. In: CLEF 2018: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 343–361. Avignon, France (September 2018)
19. 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. Lugano, Switzerland (September 2019)
20. 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 (CLPsych 2018). pp. 37–46. Association for Computational Linguistics, New Orleans, LA (2018)

21. 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 (CLPsych 2016). pp. 118–127. Association for Computational Linguistics, San Diego, CA, USA (June 2016)
22. Mohammadi, E., Amini, H., Kosseim, L.: CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych 2019). Minneapolis, Minnesota, USA (June 2019)
23. Ofoghi, B., Mann, M., Verspoor, K.: Towards early discovery of salient health threats: A social media emotion classification technique. In: Biocomputing 2016: Proceedings of the Pacific Symposium. pp. 504–515. Kohala Coast, Hawaii (January 2016)
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS 2017 Autodiff Workshop. Long Beach, California, USA (January 2017)
25. 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**(Oct), 2825–2830 (2011)
26. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). pp. 1532–1543. Doha, Qatar (October 2014)
27. Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., Leenaars, A.: Suicide note classification using natural language processing: A content analysis. Biomedical informatics insights **3**, BII–S4706 (2010)
28. 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 (CLPsych 2017). pp. 58–65 (2017)
29. Socher, R., Lin, C.C.Y., Ng, A., Manning, C.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011). pp. 129–136. Bellevue, Washington, USA (June 2011)
30. Struik, L.L., Baskerville, N.B.: The role of facebook in crush the crave, a mobile-and social media-based smoking cessation intervention: qualitative framework analysis of posts. Journal of medical Internet Research **16**(7) (2014)
31. Thompson, J.J., Leung, B.H., Blair, M.R., Taboada, M.: Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. Knowledge-Based Systems **137**, 149–162 (December 2017)
32. 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: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. Avignon, France (September 2018)
33. Yang, C.C., Jiang, L., Yang, H., Tang, X.: Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012). Beijing, China (August 2012)
34. Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., Power, R.: Using social media to enhance emergency situation awareness. In: Twenty-Fourth Inter-

national Joint Conference on Artificial Intelligence (IJCAI 2015). Buenos Aires, Argentina (July 2015)
35. Zirikly, A., Resnik, P., Uzuner, Ö., 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: From Keyboard to Clinic (CLPsych 2019). Minneapolis, Minnesota, USA (June 2019)