Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia

FHDO Biomedical Computer Science Group (BCSG)

Marcel Trotzek[1], Sven Koitka[1,2,3], and Christoph M. Friedrich[1,4]

[1] University of Applied Sciences and Arts Dortmund (FHDO)
Department of Computer Science
Emil-Figge-Str. 42, 44227 Dortmund, Germany
mtrotzek@stud.fh-dortmund.de, sven.koitka@fh-dortmund.de, and
christoph.friedrich@fh-dortmund.de

[2] TU Dortmund University Department of Computer Science, Germany

[3] Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany

[4] Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany

Abstract. Developing methods for the early detection of mental disorders like depression and anorexia based on written text has become an important aspect with the rise of social media platforms. The CLEF 2018 eRisk shared task consists of two subtasks focussed on the detection of these two disorders and *FHDO Biomedical Computer Science Group* (BCSG) has submitted results obtained from four machine learning models as well as from a final late fusion ensemble. This paper describes these models based on user-level linguistic metadata, Bags of Words (BoW), neural word embeddings, and Convolutional Neural Networks (CNN). BCSG has achieved top performance according to $ERDE_{50}$ and F_1 score in both subtasks.

Keywords: depression, early detection, linguistic metadata, convolutional neural networks, word embeddings

1 Introduction

This paper describes the participation of *FHDO Biomedical Computer Science Group* (BCSG) at the *Conference and Labs of the Evaluation Forum* (CLEF) 2018 eRisk task for early detection of depression and anorexia [11, 13]. BCSG submitted results obtained from four different models and a late fusion ensemble of three of these models. These models as well as the findings concerning the dataset are described in this paper and an outlook on possible improvements and future research is given. The work described in this paper is based on this team's previous participation in the eRisk 2017 pilot task for early detection of depression [27] and on further research based on the same dataset [28].

2 Related Work

Studies concerning the effect of mental state on the language used by a person have already shown various connections, beginning with observations of more frequent uses of first personal singular pronouns in spoken language of depression patients [4, 29]. More recent studies found, for example, an elevated use of the word "I" in particular and more negative emotion words in essays by depressed college students [19], more verbs in past tense and pronouns in general spoken by Russian depression patients [25], and a more frequent use of absolutist words (e.g. absolutely, completely, every, nothing) in forums related to depression, anxiety, or suicidal ideation than in unrelated forums or forums about asthma, diabetes, and cancer [2].

Results like these have lead to the development of tools that allow researchers and therapists to evaluate written texts with a focus on the author's mental state. One such tool is the Linguistic Inquiry and Word Count (LIWC) software [26], which calculates a total of 93 features for any given text document based on a dictionary. Similarly, Differential Language Analysis Toolkit (DLATK) [22] was published as an open-source Python library for text analysis with a focus on psychology, health, and social aspects.

First results in the area of early detection of depression based on written social media texts have been reported as part of the eRisk 2017 pilot task [12]. Similar research without the early detection aspect has previously been done, for example, at the CLPsych shared task for detection of depression and PTSD on Twitter [5]. In the same domain as this task, data from reddit.com has recently been utilized to successfully detect messages concerning anxiety [24].

3 Datasets and Tasks

Similar to the task in 2017, the datasets of both subtasks consist of messages obtained from the social media platform reddit.com. The training data of the depression subtask is equivalent to the full training and test data of this previous task, while the anorexia subtask is based on completely new messages. An especially interesting aspect of reddit is that it allows users to create communities with specific topics called *subreddits*. There exists a wide variety of these communities, also including very active ones from a depression detection perspective, like /r/depression[5], which is mainly used by people struggling with depression.

The messages contained in both datasets can consist of a separate *title* and *text* field depending on the type of message: Users can post content in terms of links or images (only *title*, link or image not included), text content (*title* and optional *text*), or as comment on another message (only *text*). Some messages in both datasets also include no *text* or *title* and can therefore be discarded. The number of documents per user ranges between 10 and 2000. In every week of the test phase, a chunk of 10% of each user's messages is supplied to the

[5] http://www.reddit.com/r/depression, Accessed on 2018-04-02

participants in chronological order, resulting in 1 to 200 documents per user each week. In both subtasks there are exceptions of this general rule because one anorexia training user (subject2167 of the control group), three depression test users (subject5161, subject5301, and subject8719), and two anorexia test users (subject4169 and subject7483) do not have any messages in the final week, resulting in only 9 messages of these users.

Table 1 displays the main characteristics of the two datasets. The average amount of characters and unigrams per document was calculated based on a concatenation of the *text* and *title* field. To calculate the number of unigrams, the same preprocessing and tokenization as described in sections 4.3 and 4.4 was utilized, retaining only words that occur in the writings of at least two users.

Table 1. Characteristics of the training and test datasets for both subtasks.

	Depression		Anorexia	
	Training	Test	Training	Test
Users	887	820	152	320
Positive/Negative	135/752	79/741	20/132	41/279
Documents	531,394	544,447	84,834	168,507
Comments (empty title)	367,439	366,845	61,201	130,631
One-liners (empty text)	141,849	147,197	15,768	27,228
Empty documents	91	219	39	90
Avg. documents per user	599.09	663.96	558.12	526.58
Avg. characters per doc.	174.54	197.47	178.11	171.36
Avg. unigrams per doc.	30.83	34.53	31.59	30.95
Unique unigrams	85,558	94,569	31,128	45,727

3.1 Hand-crafted User Features

The participation of this team in the eRisk 2017 pilot task was based on a set of user-level linguistic metadata features that were used as additional input for every model. In this second eRisk shared task, only one submitted model (see section 4.1) and the final late fusion ensemble (see section 4.5) use metadata features. All text based features have again been calculated based on a concatenation of the *text* and *title* field of each message. Still, this includes the same set of features described in the previous working notes paper [27] and an additional set of ten features obtained from the Linguistic Inquiry and Word Count (LIWC) [26] software. These LIWC features have been chosen based on their correlation with the class label in the depression subtask training data. Another addition to the original feature set is the average length of the *title* field that was also not used in 2017.

Figure 1 illustrates the correlation matrix of the complete metadata feature set and includes the class label information to indicate the relevance of each feature. Although some features—especially the pronoun counts—seem redundant at first sight, all of the original features are preserved as they are based on a Part of Speech (POS) tagging using the Python NLTK framework[6] while LIWC features are based on a lexicon that also includes abbreviations or common misspellings. Most of the described features are averaged over all documents per user to obtain the final metadata feature vector, except for the counts of specific phrases like medication names or mentioned diagnoses which are summed. Finally, all averaged features are standardized to have unit variance and a mean of 0 and the summed features are converted to flags with a value of 1 for users that have used such a phrase in any document and -1 otherwise.

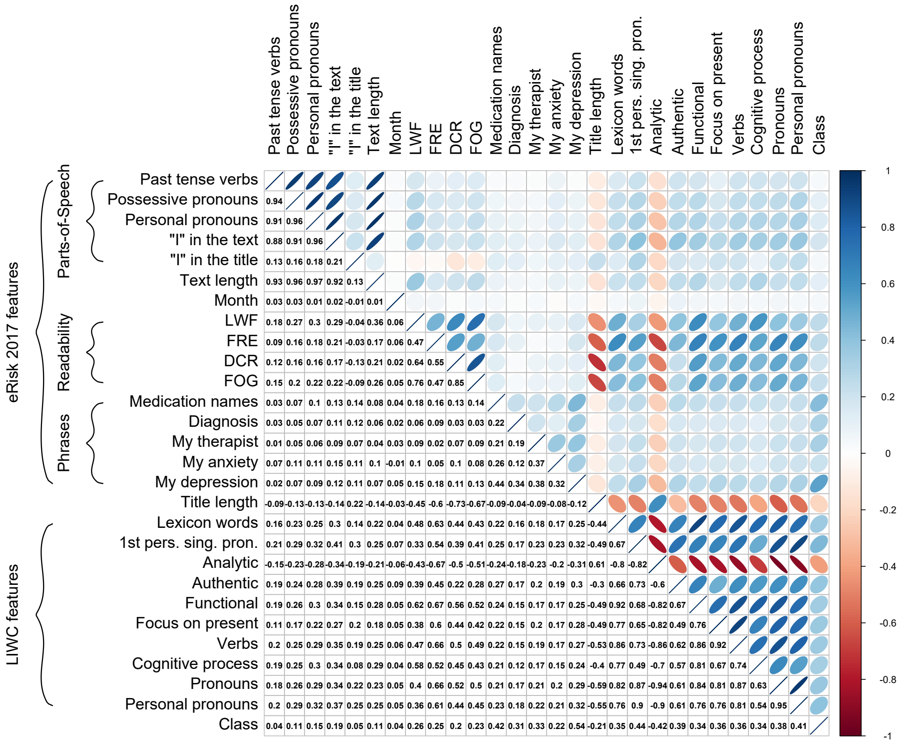

Fig. 1. Correlation matrix of all user features including the class information (non-depressed/depressed) based on the depression subtask training data. This plot is best viewed in electronic form.

[6] http://www.nltk.org/book/ch05.html, accessed on 2018-04-02

4 Chosen Models

This section describes the five models that have been used to classify the test users of both subtasks. The models for both tasks are completely identical, use the same set of metadata features, and only vary slightly in their prediction thresholds as described below. In comparison to this team's participation in the eRisk 2017 pilot task, the prediction thresholds were simplified: For each model, only a single prediction threshold value was chosen based on cross-validation on the training data to indicate whether a subject is classified as *depressed*. The number of documents already processed for a user is not used anymore as the new models are less prone to predict many false positives after processing only few documents. In addition, *non-depressed* predictions are now only submitted in the final week because early prediction of these cases has no effect on the score and later writings might still identify them as *depressed*. Selecting viable prediction thresholds is difficult as a balanced result according to both $ERDE_o$ and F_1 is often hard to achieve. The goal for this participation was to use rather low thresholds to find *depressed* cases as early as possible without generating too many false positives.

In contrast to the previous participation of this team, only the first model and the final ensemble utilize the updated set of user metadata features described in section 3.1. The bag of words model, which achieved the best overall F_1 as well as second best $ERDE_5$ and $ERDE_{50}$ score in the previous task [12], is reused with and without metadata features. The Recurrent Neural Network (RNN) using a Long Short Term Memory (LSTM) layer was not evaluated again and instead replaced with a Convolutional Neural Network (CNN). This decision was based on further research using the eRisk 2017 dataset [28], which showed that the CNN model was able to outperform results of the LSTM models and also easier to configure and less prone to overfitting.

4.1 Bag of Words Metadata Ensemble - BCSGA

The first model is mostly equivalent to the first model used in this team's participation in eRisk 2017, except for the extended set of metadata features. It utilizes an ensemble of *Bag of Words* (BoW) classifiers with different term weightings and n-grams that are calculated on a user basis by first concatenating all documents (text and title) of a user. The term weighting for bags of words can generally be split into three components: a *term frequency component* or local weight, a *document frequency component* or global weight, and a *normalization component* [21]. A general term weighting scheme can therefore be given as [30]:

$$t_{t,d} = l_{t,d} \cdot g_t \cdot n_d \ , \tag{1}$$

where $t_{t,d}$ is the calculated weight for term t in document d, $l_{t,d}$ is the local weight of term t in document d, g_t is the global weight of term t for all documents, and n_d is the normalization factor for document d. A common example would be using the *term frequency* (tf) as local weight and the *inverse document frequency* (idf) as global weight, resulting in tf-*idf* weighting [21].

All ensemble models use l2-norm for n_d but varying local and global weights. The first one uses a combination of uni-, bi-, tri-, and 4-grams obtained from the training data. To build this first BoW, the 200,000 $\{1, 2, 3, 4\}$-grams with the highest *Information Gain* (IG) are selected, given by [14, p. 272]:

$$I(U, C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U = e_t, C = e_c) \cdot log_2 \left(\frac{P(U = e_t, C = e_c)}{P(U = e_t) \cdot P(C = e_c)} \right) \quad , \tag{2}$$

with the random variable U taking values $e_t = 1$ (the document contains term t) and $e_t = 0$ (the document does not contain term t) and the random variable C taking values $e_c = 1$ (the document is in class c) and $e_c = 0$ (the document is not in class c). The raw term frequency of the resulting n-grams is used as local weight, while their IG-score is used as global weight. The second BoW utilizes a modified version of tf, namely *augmented term frequency* (atf) [30], multiplied by idf:

$$atf\text{-}idf(t, d) = \left(a + (1 - a) \frac{tf_t}{\max(tf)} \right) \cdot \log \frac{n_d}{df(d, t)} \quad , \tag{3}$$

with $\max(tf)$ being the maximum frequency of any term in the document, the total number of documents n_d, and the smoothing parameter a, which is set to 0.3 for this model. This BoW, as well as the third one, contains all unigrams of the training corpus. The local weight of the third model consists of the *logarithmic term frequency* ($logtf$) [16] and the global weight is given by *relevance frequency* (rf) [9], which can be combined as:

$$logtf\text{-}rf(t, d) = (1 + \log(tf)) \cdot \log_2 \left(2 + \frac{df_{t,+}}{\max(1, df_{t,-})} \right) \quad , \tag{4}$$

where $df_{t,+}$ and $df_{t,-}$ is the number of documents in the *depressed*/*non-depressed* class that contain the term t. The final model of this ensemble uses the hand-crafted user features described in section 3.1.

All three bags of words and the hand-crafted features were each used as input for a separate logistic regression classifier. Due to the imbalanced class distribution, a modified class weight was used for these classifiers similar to the original task paper [11] to increase the cost of false negatives. It was calculated for the *non-depressed* class as $1/(1+w)$ and for the *depressed* class as $w/(1+w)$, with $w = 2$ for all four models. The final output probabilities were calculated as unweighted mean of all four logistic regression probabilities. Each week and for both tasks, this ensemble predicted any user with a probability above or equal to 0.4 as *depressed*, while in the final week all users with a probability less than 0.4 were predicted as *non-depressed*.

4.2 Bag of Words Ensemble - BCSGB

The second model is similar to the first one, but it only includes the three bags of words in the ensemble and disregards the metadata features. Again, for the depression subtask any test subject with a probability of at least 0.4 was predicted as *depressed*, while users with a probability below 0.4 were predicted as *non-depressed* in the final week. The prediction threshold for the anorexia subtask was set to 0.3 in this case.

4.3 CNN with GloVe Embeddings - BCSGC

The third model consists of a Convolutional Neural Network (CNN) [10], which have previously been utilized by many recent studies to achieve outstanding results especially in the area of image classification and are generally viable for data with a grid-like structure [6]. The implementation has been done based on Tensorflow [1] and the input of this CNN is based on GloVe [18] word embeddings: A 50-dimensional set of word embeddings pre-trained on Wikipedia and News[7] is used to produce a matrix of word vectors for the first 100 words of each document in the dataset. Prior to this vectorization, the documents are preprocessed and tokenized in a way that preserves, for example, emoticons, punctuation, words including special characters, and generally all tokens that occur in the documents of at least two users. Zero-padding is used for documents with less than 100 words. Each document is therefore represented by a 100×50 matrix and is classified independently. Since the number of words per document in the training data ranges between 1 (when ignoring the empty documents) and 6,487 but has a mean of 34.58 according to the tokenization done for this work, the limitation to 100 words (or even fewer to minimize the necessary zero-padding) is viable.

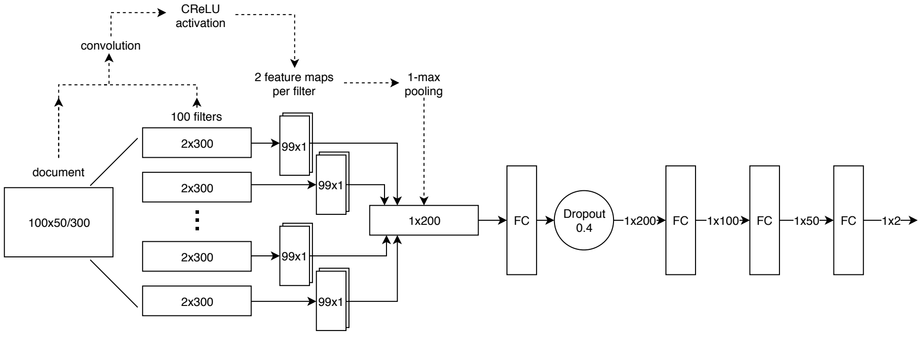

Fig. 2. Architecture of the convolutional neural network used for the models BCSGC and BCSGD (with 300 instead of 50 dimensional word vectors) [28].

The text classification network architecture used for this work is displayed in Figure 2, which shows the use of 300 dimensional word vectors (and therefore 100×300 documents) as used for the next model BCSGD. It is similar to the

[7] http://nlp.stanford.edu/projects/glove, Accessed on 2018-03-30

one-layer CNN for sentence classification described by Zhang and Wallace [31] and consists of only a single convolutional layer, 100 filters with an equal height of 2 and a width corresponding to the word embedding dimensions, and uses 1-max pooling to extract a single value from each filter. Due to the usage of Concatenated Rectified Linear Units (CReLU) [23] activation, this finally results in a 200-dimensional vector per document that is propagated through four fully connected layers, of which the first applies dropout to its output and the final one applies softmax. The training steps of this and the following CNN model utilized Adam [8] to minimize the cross-entropy loss. Both models were trained using a learning rate of $1e{-}4$ and a batch size of 10.000 documents. BCSGC was trained for 30 epochs.

To obtain a final prediction per user, the 98th percentile of the outputs from all the user's documents is calculated. This ensures that even *depressed* users that have very few documents with a high probability can be correctly predicted. For both subtasks, any subject with a final probability of at least 0.4 was predicted as *depressed* in each week, while probabilities below 0.4 again resulted in a *non-depressed* prediction in the final week.

4.4 CNN with fastText Embeddings - BCSGD

The second CNN model is based on the same architecture as the previous one but utilizes 300-dimensional fastText [7, 3, 15] word embeddings. To evaluate word vectors that are more related to the domain of reddit messages or social media in general, a new fastText model was trained specifically for this task. A dataset of all 1.7 billion reddit comments written between October 2007 and May 2015[8] was used as training corpus for this model and preprocessed similar to the description in section 4.3 but without removing infrequent words yet. In addition to this, any references to reddit users (in the form of /u/<username>) were replaced by a generic phrase "ref_user" to prevent any connections to actual users in the resulting word embeddings. Similarly, any reference to a subreddit (in the form of /r/<subreddit>) was replaced by the phrase "ref_subreddit_<subreddit>" to be able to learn a vector representation of them as well that can be regarded as their topic. No stemming or stopword removal of any kind was done and messages in other languages than English were removed based on stopword counts. The final corpus of 1.37 billion reddit comments was used to train 6 million word vectors of words that occur at least five times in the corpus. Additional details about this model and the utilized CNNs can be found in the corresponding paper [28].

Similar to the previous CNN model, the resulting 100 × 300 matrix of word embeddings obtained for each document was classified separately and the 98th percentile of the outputs was used as output for the corresponding user. This model was trained for 25 epochs using the same parameters as BCSGC. The prediction threshold for *depressed* predictions was set to 0.7 for both tasks,

[8] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/, Accessed on 2018-03-30

leading to a *non-depressed* prediction for probabilities below 0.7 in the final week.

4.5 CNN and Bag of Words Metadata Ensemble - BCSGE

The final model consists of a simple late fusion ensemble that has been calculated as the unweighted mean of the outputs obtained from models BCSGA, BCSGC, and BCSGD - the bag of words including metadata and the two CNN models. Although these outputs have not been calibrated (e.g. by using Platt scaling [17]) and can therefore not be seen as directly comparable probabilities, previous experiments [28] have shown that such an ensemble was able to improve the results of the separate models. Again, a prediction threshold of at least 0.4 was used for the depression detection subtask, while a threshold of 0.5 was utilized for the anorexia subtask.

5 Results

Before examining the results of the described models in the two subtasks, it is necessary to analyze the utilized $ERDE_o$ metric for early detection systems. Since this metric is based on the absolute number of documents read per user before a true positive prediction, but these documents have to be read in ten equally sized chunks, the score is highly dependent on the number of documents available per user. Because at least 10% of each user's documents have to be read by all participants, it is impossible to predict some users correctly depending on the parameter o that describes after how many documents the penalty for late predictions grows. This fact has already been described in more detail in another paper [28].

Table 2 displays the best $ERDE_5$ and $ERDE_{50}$ scores that are possible for the test data of the depression and anorexia subtask. These results are based on a perfect prediction in the first week of the tasks. As described in the above-mentioned paper, only test users with less than 100 documents (less than 10 per chunk) have any effect on the $ERDE_5$ score. This means that only predicting 26 of the 79 *depressed* test users correctly in the first week and ignoring all others still leads to an $ERDE_5$ score of 7.78 (F_1 = 0.50), while predicting only 12 of the 41 *anorexia* users in the first week also leads to an $ERDE_5$ score of 10.23 (F_1 = 0.45). $ERDE_5$ alone, without the additional F_1 score, is therefore hard to interpret.

To examine the weekly predictions obtained from the described models, Figures 3 and 4 show the cumulative number of positive predictions for the two subtasks and also visualize the proportion of true positives. For the depression subtask, this shows that the ensemble indeed lead to the most true positives but also many false positives. BCSGD seems to perform worse at first sight but indeed achieved a good balance between true and false positives because of its higher prediction threshold. As the comparison of both figures shows, the

Table 2. Best possible $ERDE_o$ scores of both subtasks based on a perfect prediction in the first week.

	Depression	Anorexia
$ERDE_5$	7.78	10.23
$ERDE_{50}$	3.79	4.05

anorexia subtask was much easier using the same models and that it was possible to detect nearly all positive samples without too many false positives. Both examinations show a steady progression over the ten weeks for all models.

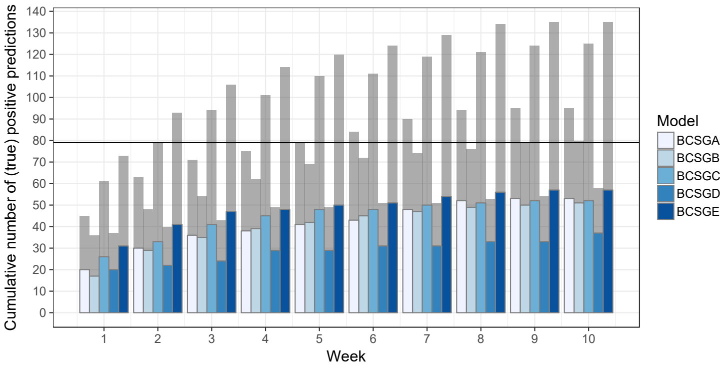

Fig. 3. Cumulative number of *depressed* predictions (blue plus gray bars) and proportion of true positives (blue bars only) per model after each week of the depression subtask. A horizontal line marks the 79 *depressed* samples in the test data.

Tables 3 and 4 show the official results [13] of BCSG's models for both subtasks and also include the alternative early detection scores $F_{latency}$ [20] and $ERDE_o^{\%}$ [28]. According to the suggestion in the paper, $F_{latency}$ was calculated using a value for the parameter p that fits the true positive cost function $P_{latency}$ to return a cost of 0.5 for the median number of documents of the positive test users. This results in a value of $p = 0.0051$ for the depression subtask (median of 216 documents per *depressed* test user) and $p = 0.0042$ for the anorexia subtask (median of 260 documents per *anorexia* test user). In contrast to the standard $ERDE_o$ score, $ERDE_o^{\%}$ is calculated based on the percentage of read documents per user and is therefore easier to interpret in a chunk-based task. Additional results by other teams have been added to these tables to include at least the best two results obtained for each score.

While the direct comparison of BCSGA (bags of words with linguistic metadata) and BCSGB (bags of words only) shows that the metadata features result in more positive predictions, the actual amount of true positives was only better for the depression subtask and resulted in a better $ERDE_5$ score but worse

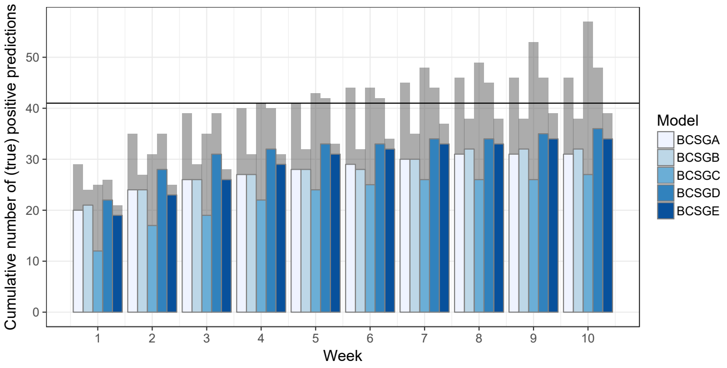

Fig. 4. Cumulative number of *anorexia* predictions (blue plus gray bars) and proportion of true positives (blue bars only) per model after each week of the anorexia subtask. A horizontal line marks the 41 *anorexia* samples in the test data.

Table 3. Best results of the depression subtask based on the official evaluation and the alternative metrics $F_{latency}$ and $ERDE_o^{\%}$. The models have been chosen to show at least the two best results achieved in each score.

Model	$ERDE_5$	$ERDE_{50}$	F_1	$F_{latency}$	$ERDE_{20}^{\%}$	$ERDE_{50}^{\%}$
BCSGA	9.21	6.68	0.61	0.47	7.08	5.31
BCSGB	9.50	**6.44**	**0.64**	**0.52**	7.17	5.04
BCSGC	9.58	6.96	0.51	0.41	6.89	4.82
BCSGD	9.46	7.08	0.54	0.41	7.32	6.34
BCSGE	9.52	6.49	0.53	0.43	**6.16**	4.57
LIIRB	10.03	7.09	0.48	0.39	10.66	5.05
UNSLA	**8.78**	7.39	0.38	0.25	7.45	6.96
UNSLD	10.68	7.84	0.45	0.37	6.23	**4.52**
UNSLE	9.86	7.60	0.60	0.45	7.76	5.50

$ERDE_{50}$ and F_1. Similar to the task in 2017, the bag of words ensemble again obtained the best results in the depression subtask, while the CNN based on the self-trained fastText embeddings (BCSGD) and the ensemble using both the bags of words as well as the CNNs (BCSGE) achieved the best scores in the anorexia subtask. Overall, the models of BCSG achieved the second-best results in $ERDE_5$ and the best results in all other scores except for another second-best result according to $ERDE_{50}^{\%}$ in the depression subtask.

As already described, the $ERDE_o$ score and especially $ERDE_5$ should be discussed in more detail because of the fact that optimizing it can often lead to simply minimizing false positives by only predicting very few users at all. A detailed look at the results and the achieved $ERDE_5$ scores shows that, for example, in the first week of the depression subtask both UNSLA and BCSGA have predicted 45 users as depressed of which 20 were indeed true positives. Still,

Table 4. Best results of the anorexia subtask based on the official evaluation and the alternative metrics $F_{latency}$ and $ERDE_o^{\%}$. The models have been chosen to show at least the two best results achieved in each score.

Model	$ERDE_5$	$ERDE_{50}$	F_1	$F_{latency}$	$ERDE_{20}^{\%}$	$ERDE_{50}^{\%}$
BCSGA	12.17	7.98	0.71	0.64	6.54	4.82
BCSGB	11.75	6.84	0.81	0.74	6.02	4.46
BCSGC	13.63	9.64	0.55	0.47	9.48	6.83
BCSGD	12.15	**5.96**	0.81	0.75	**5.48**	**3.14**
BCSGE	11.98	6.61	**0.85**	**0.78**	6.45	3.64
LIIRA	12.78	10.47	0.71	0.57	13.05	5.55
PEIMEXB	12.41	7.79	0.64	0.57	6.86	5.61
RKMVERIA	12.17	8.63	0.67	0.59	6.76	6.76
UNSLB	**11.40**	7.82	0.61	0.54	6.84	6.53
UNSLD	12.93	9.85	0.79	0.63	9.03	6.68

the resulting $ERDE_5$ score differs drastically because the predicted users vary in the number of total documents and even though UNSLA only had five more true positives in the following nine weeks, while BCSGA already had ten more in the second week and a total of 53 in the end. Similarly, the leading model in $ERDE_5$ of the anorexia subtask, UNSLB, had 19 true positives in the first week, while BCSGD already had 22, BCSGB had 21, and BCSGA had 20. In summary, $ERDE_5$ produces highly misleading results because of the varying number of documents per user.

6 Conclusions

Again, the eRisk competition has been a challenging task concerning the early detection of mental health issues based on sequences of social media texts. The depression subtask had similar F_1 scores but much better $ERDE_o$ scores based on a test set that was nearly as large as last year's training and test set combined. The results of the anorexia subtask were surprisingly good, which probably is due to the nature of this dataset. Generally, the promising results with the test data of eRisk 2017 obtained only based on linguistic metadata [28] could not yet be confirmed in this year's tasks. As already concluded in the same paper, finding a way to successfully integrate the metadata features into the neural network models is an interesting task for future research.

The examination of the task results again shows that a discussion about a meaningful metric should be a priority in the future. Both $F_{latency}$ and $ERDE_o^{\%}$ include interesting ideas to improve the evaluation of early prediction models. $F_{latency}$ contains a cost function that grows less rapidly and already incorporates the F_1 score, which makes it more meaningful when viewed alone. $ERDE_o^{\%}$ is more viable for chunk-based shared tasks because it is calculated based on the proportion of read documents per user instead of the absolute number, which leads to results that are better interpretable than the standard $ERDE_o$. A combination of these two ideas could be a promising basis for discussions about future early detection tasks.

7 Acknowledgment

The work of Sven Koitka was partially funded by a PhD grant from University of Applied Sciences and Arts Dortmund, Germany.

References

1. Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M., Levenberg J., Monga R., Moore S., Murray D.G., Steiner B., Tucker P., Vasudevan V., Warden P., Wicke M., Yu Y., Zheng X.: TensorFlow: A System for Large-Scale Machine Learning 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI'16), pp. 265–283, Savannah, Georgia, USA (2016)
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, Prepublished January 5, 2018, DOI: 10.1177/2167702617747074 (2018)
3. Bojanowski, P.,Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135–146 (2017)
4. Bucci, W., Freedman, N.: The Language of Depression. Bulletin of the Menninger Clinic, Vol. 45(4), pp. 334–358 (1981)
5. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych 2015 Shared Task: Depression and PTSD on Twitter. Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych'15), pp. 31–39, Denver, Colorado, USA (2015)
6. Goodfellow I., Bengio Y., Courville A.: Deep Learning. MIT Press (2016)
7. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Vol. 2, pp. 427–431 (2016)
8. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, California, USA, arXiv preprint arXiv:1412.6980 (2015)
9. Lan, M., Tan, Chew L., Low, H.-B.: Proposing a New Term Weighting Scheme for Text Categorization. Proceedings of the 21st National Conference on Artifical Intelligence (AAAI-06), Vol. 6, pp. 763–768, Boston, Massachusetts, USA (2006)
10. LeCun, Y.: Generalization and Network Design Strategies. Technical Report CRG-TR-89-4, University of Toronto (1989)
11. Losada, D.E., Crestani, F.: A Test Collection for Research on Depression and Language Use. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, pp. 28–39. CLEF 2016, Évora, Portugal (2016)
12. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. Proceedings Conference and Labs of the Evaluation Forum CLEF 2017, Dublin, Ireland (2017)

13. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France (2018)
14. Manning, C.D., Raghavan, P., Schütze, H.: An Introduction to Information Retrieval. Online Edition. Cambridge University Press (2009) Available from: https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf - Accessed on 2018-04-02
15. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A.: Advances in Pre-Training Distributed Word Representations. Proceedings of the International Conference on Language Resources and Evaluation (LREC'18), Miyazaki, Japan (2018)
16. Paltoglou, G., Thelwall, M.: A Study of Information Retrieval Weighting Schemes for Sentiment Analysis. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1386–1395. Association for Computational Linguistics (2010)
17. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers, Vol. 10(3), pp. 61–74 (1999)
18. Pennington J., Richard S., Manning C.D.: GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14), ACL, pp. 1532–1543, Doha, Qatar (2014)
19. Rude, S., Gortner, E.-M., Pennebaker, J.: Language Use of Depressed and Depression-Vulnerable College Students. Cognition & Emotion, Vol. 18(8), pp. 1121–1133 (2004)
20. Sadeque, F., Xu, D., Bethard, S.: Measuring the Latency of Depression Detection in Social Media. Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'18), pp. 495–503, Los Angeles, California, USA (2018)
21. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management, Vol. 24(5), pp. 513–523 (1988)
22. Schwartz, H.A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., Eichstaedt, J.: DLATK: Differential Language Analysis Toolkit. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP'17), ACL, pp. 55–60, Copenhagen, Denmark (2017)
23. Shang, W., Sohn, K., Almeida, D., Lee, H.: Understanding and Improving Convolutional Neural Networks via Concatenated Rectified Linear Units Proceedings of The 33rd International Conference on Machine Learning, Vol. 48, pp. 2217–2225, New York City, New York, USA (2016)
24. Shen, J.H., Rudzicz, F.: Detecting Anxiety through Reddit. Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology. From Linguistic Signal to Clinical Reality (CLPsych'17), pp. 58–65, Vancouver, Canada (2017)
25. Smirnova, D., Sloeva, E., Kuvshinova, N., Krasnov, A., Romanov, D., Nosachev, G.: Language Changes as an Important Psychopathological Phenomenon of Mild Depression. European Psychiatry, Vol. 28 (2013)
26. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, Vol. 29(1), pp. 24–54 (2010)

27. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. Working Notes Conference and Labs of the Evaluation Forum CLEF 2017, Dublin, Ireland (2017) Available from: http://ceur-ws.org/Vol-1866/paper_54.pdf - Accessed on 2018-03-29
28. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. arXiv preprint arXiv:1804.07000 [cs.CL] (2018)
29. Weintraub, W.: Verbal Behavior: Adaptation and Psychopathology. Springer Publishing Company (1981)
30. Wu, H., Gu, X.: Reducing Over-Weighting in Supervised Term Weighting for Sentiment Analysis. The 25th International Conference on Computational Linguistics (COLING 2014), pp. 1322–1330, Dublin, Ireland (2014)
31. Zhang, Y., Wallace, B.: A Sensitivity Analysis of (and Practioners' Guide to) Convolutional Neural Networks for Sentence Classification. Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Asian Federation of Natural Language Processing, pp. 253–263, Taipei, Taiwan (2017)