Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences

Marcel Trotzek , Sven Koitka , and Christoph M. Friedrich , *Member, IEEE*

Abstract—Depression is ranked as the largest contributor to global disability and is also a major reason for suicide. Still, many individuals suffering from forms of depression are not treated for various reasons. Previous studies have shown that depression also has an effect on language usage and that many depressed individuals use social media platforms or the internet in general to get information or discuss their problems. This paper addresses the early detection of depression using machine learning models based on messages on a social platform. In particular, a convolutional neural network based on different word embeddings is evaluated and compared to a classification based on user-level linguistic metadata. An ensemble of both approaches is shown to achieve state-of-the-art results in a current early detection task. Furthermore, the currently popular $ERDE$ score as metric for early detection systems is examined in detail and its drawbacks in the context of shared tasks are illustrated. A slightly modified metric is proposed and compared to the original score. Finally, a new word embedding was trained on a large corpus of the same domain as the described task and is evaluated as well.

Index Terms—Depression, early detection, linguistic metadata, convolutional neural network, word embeddings

✦

1 INTRODUCTION

ACCORDING to World Health Organization (WHO) [1], more than 300 million people worldwide are suffering from depression, which equals about 4.4 percent of the global population. While forms of depression are more common among females (5.1 percent) than males (3.6 percent) and prevalence differs between regions of the world, it occurs in any age group and is not limited to any specific life situation. Depression is therefore often described to be accompanied by paradoxes, caused by a contrast between the self-image of a depressed person and the actual facts [2]. Latest results from the 2016 National Survey on Drug Use and Health in the United States [3] report that, during the year 2016, 12.8 percent of adolescents between 12 and 17 years old and 6.7 percent of adults had suffered a major depressive episode (MDE).

Precisely defining depression is not an easy task, not only because several sub-types have been described and changed in the past [4], but also because the term "being depressed" has become frequently used in everyday language. In general, depression can be described to lead to an altered mood and may also be accompanied, for example, by a negative self-image, wishes to escape or hide, vegetative changes, and a lowered overall activity level [2, p. 8]. The symptoms experienced by depressed individuals can severely impact their ability to cope with any situation in daily life and therefore differ drastically from normal mood variations that anyone experiences.

At the worst, depression can lead to suicide. WHO estimates that, in the year 2015, 788,000 people have died by suicide and that it was the second most common cause of death for people between 15 and 29 years old worldwide [1]. In Europe, self-harm was even reported as the most common cause of death in the age group between 15 and 29 and the second most common between 30 and 49, again in results obtained by WHO in 2015 [5].

Although the severity of depression is well-known, only about half of the individuals affected by any mental disorder in Europe get treated [6]. The proportion of individuals seeking treatment for mood disorders during the first year ranges between 29–52 percent in Europe, 35 percent in the USA, and only 6 percent in Nigeria or China [7]. In addition to possible personal reasons for avoiding treatment, this is often due to a limited availability of mental health care, for example in conflict regions [8]. Via a telephone survey in Germany [9],

- *M. Trotzek is with the Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund 44227, Germany. E-mail: mtrotzek@stud.fh-dortmund.de.*
- *S. Koitka is with the Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund 44227, Germany, the Department of Computer Science, TU Dortmund University, Dortmund 44227, Germany, and the Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen 45147, Germany. E-mail: sven.koitka@fh-dortmund.de.*
- *C.M. Friedrich is with the Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund 44227, Germany, and the Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Essen 45130, Germany. E-mail: christoph.friedrich@fh-dortmund.de.*

researchers found out that shame and self-stigmatization seem to be much stronger reasons to not seek psychiatric help than actual perceived stigma and negative reactions of others. They further speculate that the fear of discrimination might be relatively unimportant in their study because people hope to keep their psychiatric treatment secret. Another study amongst people with severe mental illness in Washington D. C. showed that stigma and discrimination indeed exist, while they are not "commonly *experienced* problems" but rather "perceived as omnipresent *potential* problems" [10, p. 1].

While depression and other mental illnesses may lead to social withdrawal and isolation, it was found that social media platforms are indeed increasingly used by affected individuals to connect with others, share experiences, and support each other [11], [12]. Based on these findings, peer-to-peer communities on social media can be able to challenge stigma, increase the likelihood to seek professional help, and directly offer help online to people with mental illness [13]. A similar study in the USA [14] came to the conclusion that internet users with stigmatized illnesses like depression or urinary incontinence are more likely to use online resources for health-related information and for communication about their illness than people with another chronic illness. All this emphasizes the importance of research toward ways to assist depressed individuals on social media platforms and on the internet in general.

This paper is therefore focused on ways to classify indications of depression in written texts as early as possible based on machine learning methods. The work presented in this paper is structured as follows: Section 2 gives an overview of related work concerning depression, its influence on language, and natural language processing methods. Section 3 describes the dataset used in this work, analyzes the evaluation metric of the corresponding task, and proposes an alternative. Section 4 introduces the user-based metadata features used for classification, while Section 5 describes the neural network models utilized for this task. Section 6 contains an experimental evaluation of these models and compares them to published results. Finally, Section 7 concludes this work and summarizes the results.

2 RELATED WORK

This section describes the context of this work based on previous research concerning depression and its effects on language. Since social media research in general and health research in particular require ethical considerations, an overview of the current ethical discussion in the field of natural language processing is given. Finally, the practical basis of this work is described by investigating previous and current work in text classification using machine learning.

2.1 Depression and Language

Previous studies have already shown that depression also has an effect on the language used by affected individuals. For example, a more frequent use of first person singular pronouns in spoken language was first observed in 1981 [15], [16]. An examination of essays written by depressed, formerly-depressed, and non-depressed college students at University of Texas [17] confirmed an elevated use of the word "I" in particular and also found more negative emotion words in the depressed group. Similarly, a Russian speech study [18] found a more frequent use of all pronouns and verbs in past tense among depression patients. A recent study based on English forum posts [19] observed an elevated use of absolutist words (e.g., absolutely, completely, every, nothing[1]) within forums related to depression, anxiety, and suicidal ideation than within completely unrelated forums as well as ones about asthma, diabetes, or cancer.

The knowledge that language can be an indicator of an individual's psychological state has, for example, lead to the development of the Linguistic Inquiry and Word Count (LIWC) software [20], [21]. By utilizing a comprehensive dictionary, it allows researchers to evaluate written texts in several categories based on word counts. A more detailed description of LIWC and its features is given in Section 4. With a similar purpose, Differential Language Analysis Toolkit (DLATK) [22], an open-source Python library, was created for text analysis with a psychological, health, or social focus.

2.2 Ethical Perspective

Driven by the growing availability of data, for example through social media, and the technological advances that allow researchers to work with this data, ethical considerations are becoming more and more important in the field of Natural Language Processing (NLP). Based on these developments, NLP has changed from being mostly focussed on improving linguistic analysis towards actually having an impact on individuals based on their writings. Still, a proper discussion about ethics in NLP has only been started in 2016 by Hovy and Spruit [23]. Although Institutional Review Boards (IRBs) have been well-established to enforce ethical guidelines on experiments that directly involve human subjects, the authors note that NLP and data sciences in general have not constructed such guidelines. They further argue that language "is a *proxy for human behavior*, and a strong signal of individual characteristics" and that, in addition, "the texts we use in NLP carry *latent* information about the author and situation" [23, p. 592]. On top of this direct connection to the individual, they also describe the social impact of NLP research [23, pp. 593–594]. A demographic bias in the selection of training texts can lead to the *exclusion* of specific groups, *overgeneralization* based on false positives can have serious consequences depending on the task, and research results can potentially cause or confirm biases and ultimately discrimination by *topic overexposure*. Even if all these factors are considered, they conclude that *dual-use problems* can exist for any research if results are used in a different way than originally intended. The same applies to pre-trained machine learning models that get published and could theoretically be used in unintended ways.

These discussions about ethics in NLP have lead to the First Workshop on Ethics in Natural Language Processing[2] during the conference of the European Chapter of the Association for Computational Linguistics in 2017 (EACL 2017). Some interesting results of this workshop include, for example, a proposed process to make NLP research "ethical by

---

1. A full list of words is available as Table S2 from http://journals.sagepub.com/doi/suppl/10.1177/2167702617747074, accessed on Feb. 14, 2018.

2. http://www.ethicsinnlp.org/, accessed on Feb. 14, 2018.

design" [24] by installing an Ethics Review Board (ERB) in research organizations that has to approve or veto all steps during research, development, and deployment. Specifically for health research in social media, guidelines for ethical research have been proposed [25]. They include obtaining consent from users whenever possible, carefully considering the consequences of any interactions with users or modifications of the user experience, protecting the data during research and when sharing it with other researchers, and de-identifying users during analysis, presentation, and when linking data from several platforms. From another perspective, there are also ethical considerations to keep in mind for NLP shared tasks and shared tasks in general [26]. The competitive nature of such tasks may lead researchers to be secretive about their systems and methods, ethical concerns may be overlooked, and conflicts of interest may arise if organizers themselves participate in a task.

While most discussions about health research in social media focus on the important theoretical groundwork to establish guidelines, there has also been a qualitative study using focus group interviews with 16 depressed and 10 non-depressed participants [27] to investigate their opinion about population-level mental health monitoring on Twitter. First, participants of this study were generally aware of the fact that their Twitter messages are public, but showed misconceptions about how access to them could be limited by deleting them, by limitations of the user interface, or by the sheer amount of messages on the platform. While the participants mainly accepted aggregated depression monitoring based on Twitter, some still found it "creepy" and a particular participant stated: "The fact that if it was an algorithm, and they were looking like, 'Hey, we think you're feeling low right now.' I feel like it might make me feel even more low." [27, p. 6] Similar to this statement, participants were concerned about the possibility to use population-based data to identify specific individuals, while others had the opinion that "pinpointing individuals could help them access much-needed mental health services by paying attention to cues that friends may ignore" [27, p. 7]. In general, participants supported the idea to use social media data as an additional source for professional therapists.

2.3 Natural Language Processing

The work described in this paper belongs to the area of Natural Language Processing [28] and text classification in particular. The origins of text classification tasks can be found in early research to automatically categorize documents based on statistical analysis of specific clue words in 1961 [29]. Later, similar research goals lead to rule-based text classification systems like CONSTRUE in 1990 [30] and finally the field began to shift more and more to machine learning algorithms around the year 2000 [31], [32]. In addition to text categorization, machine learning was also a driving force in other text-based tasks like sentiment analysis, which is focussed on extracting opinions and sentiment from text documents [33]. It was first used in combination with machine learning to find positive or negative opinions in movie reviews [34] and was then extended to other review domains [35], as well as completely different areas like social media monitoring and general analysis of consumer attitudes [33].

More recently, deep learning has been utilized for text classification [36], [37] in addition to its more common usages in image classification. State-of-the-art results in several text-based tasks could, for example, be achieved by transfer learning methods like Universal Language Model Fine-tuning (ULMFit) [38] and the Google research project Bidirectional Encoder Representations from Transformers (BERT) [39] for the training of language representations, which includes ULMFit and several other methods. The code of BERT and several pre-trained models are also available on GitHub.[3]

Based on these developments, research evolved to text classification tasks that extract more than just opinions from documents: Especially the availability of social media messages enabled researchers to extract population-based health information that made it possible to track diseases, symptoms, and medications [40]. More specifically, Twitter messages were used for population level tracking of depression [41], detection of depression [42], [43], bipolar disorder [44], and post traumatic stress disorder (PTSD) [45] for individuals. Depression detection from text documents in particular has become an increasingly important research area, with interesting methods and results reported for Twitter, Facebook, and forum posts [46]. To directly help depression patients, systems like Psychologist in a Pocket [47], [48], an Android smartphone app, are being developed: Users of this app can choose specific text inputs on their device that should be monitored (e.g., social media posts, mails, or text messages) to be informed about possibly alarming mood changes that they themselves might overlook. By installing an additional plugin, data can be shared with a third party, for example a therapist, and is otherwise password secured and only saved locally.

In addition to text-based depression detection, the second sub-task of the work described in this paper can be found in the area of early detection. Early detection based on text documents can be seen to originate from the idea of sequential reading to allow predictions based on as few documents as possible [49]. An approach using a modified naïve Bayes classifier was shown to be viable for text categorization and sexual predator detection with partial information [50]. Other interesting use cases of early detection applied in practice have been found in the detection of early signs of epidemics [51] or rumors [52] from social media messages.

The fields of depression detection and early detection were first combined by the publication of a dataset for early detection of depression in reddit messages [53] and research using this dataset was driven by the Conference and Labs of the Evaluation Forum (CLEF) 2017 conference[4] workshop on early risk detection on the internet [54], [55]. As this task and dataset are also utilized in this paper, further details can be found in Section 3. During the workshop in 2017, interesting results could be obtained using combinations of Information Retrieval (IR) and supervised learning based on bag of words and dictionaries [56], a two-step classification based first on posts and then on users [57], purely user-based features and random forests [58], lexicon word counts and medial concepts using Support Vector Machines (SVM)

---

3. https://github.com/google-research/bert, accessed on Nov. 24, 2018.
4. http://clef2017.clef-initiative.eu/, accessed on Feb. 18, 2018.

TABLE 1
Main Statistics of the eRisk 2017 Dataset

	Train		Test	
	Depressed	*Control*	*Depressed*	*Control*
Users	83	403	52	349
Messages	30,851	264,172	18,706	217,665
Links/Title only	2,768	81,474	973	56,543
Title + Text	2,143	9,907	955	9,192
Comments	25,939	172,746	16,776	151,887
Empty messages	1	45	2	43
Avg. msgs. per user	371.7	655.5	359.7	623.7

Adapted from [54] and extended.

or Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU) [59], and graph models [60]. The Temporal Variation of Terms (TVT) model for early detection, based on the variation of vocabulary over time, was proposed [61] and successfully evaluated [62]. The authors of this paper participated in the task by using models that combined user-based linguistic metadata with bag of words, document embeddings, and RNNs using Long Short Term Memory (LSTM) layers [63]. Results from this task are used to evaluate the experiments in Section 6.

Similar text classification research in a psychological context has been conducted at the CLPsych conferences[5] of the past years. In 2016 and 2017, for example, the conference presented a shared task [64], [65] that challenged participants to prioritize posts in an online peer-support forum to tell moderators how urgently a message needs their attention. The CLPsych shared task in 2018 [66] focused on an even more notable approach to early detection: Based on essays written by 11-year-olds, participants had to predict the current as well as future psychological state of the author at specific times in their life.

3 DATASET OVERVIEW

This section gives an overview of the dataset used for the experiments described in this paper and its main characteristics. It also details the corresponding task and the evaluation criteria.

3.1 Dataset

The dataset utilized in all experiments for this paper was first described in 2016 for research on depression and language use [53] and then finally published as part of the CLEF 2017 conference eRisk pilot task on early detection of depression [54]. It contains chronological sequences of posts and comments from reddit.com, collected for a total of 135 depressed users and a random control group of 752 users. Depressed users were identified by searching for posts that clearly mention a diagnosis (e.g., "I was diagnosed with depression"). Since there is no way to validate these statements and no further investigation of the users was possible, there could theoretically be non-depressed individuals in this group but also depressed ones in the control group. Any occurrence of user names has been replaced by an ID like *train_subject_1* to anonymize users. The number of

messages collected for each user ranges from 10 to 2,000 due to API limitations and the fact that some of them have posted very rarely. The dataset has been split into a training and test set as displayed in Table 1.

Each message in the dataset may consist of a *title*, *text*, or both, depending on its type: Users on reddit are able to post content in terms of an image or URL (*title* only), as text content (*title* and optional *text*), or as comment on another message (*text* only). A total of 91 messages in the dataset are completely empty and can therefore be discarded. Since deleted messages are normally exchanged with the text "[deleted]", these seem to be caused by a fault in reddit, the API, or the preprocessing before publishing the dataset.

In addition, each message also contains a *date* attribute with the timestamp of when the user has published it exact to the second. Since the reddit API returns all timestamps in UTC (or the local timezone of the reddit server[6]), these timestamps can primarily be used to sort messages and search for time patterns of a single user. Comparing timestamps between different users would most likely give misleading results because their actual timezone is unknown and they could live anywhere in the world.

Since users for the control group were collected by selecting users that had posted recently when the dataset was collected, instead of using a distribution over time similar to the depressed users, the timestamps also contain a hidden feature that could be exploited: When using the time of the latest post per user (in seconds since epoch) as only input for a logistic regression, this single feature was enough to obtain an F_1 score of 0.78 on the test data. This feature could easily be used as soon as the last data chunk (see Section 3.2) is available. As this is clearly not intended and not in the interest of this task, all models created for this paper completely discard the timestamp information and a detailed analysis of this fact has been sent to the organizers of eRisk to prevent this in future tasks.[7]

3.2 Task and Evaluation Criteria

The given dataset was explicitly published for research toward early detection of depression within the previously described eRisk task. To measure this criterion, the data was also split into ten chunks by the organizers, containing 10 percent of each user's messages in chronological order. During the test phase of the eRisk task, a single chunk of data was published each week, starting with the oldest messages of the users. Participants then had the possibility to classify a user as depressed, non-depressed, or delay the decision to see additional data in the next week. Submitted predictions were final and could not be reversed later. In the last week, a prediction had to be given for every user. In addition to the correct and wrong predictions, evaluations could therefore also take into account how many messages participants had seen for each user before giving a prediction. This information can be utilized by the organizers' early risk detection error (ERDE) measure for early detection systems that was defined in their dataset paper as well [53, pp. 7–8]: With a binary

---

5. http://clpsych.org/, accessed on Nov. 24, 2018.

6. https://github.com/praw-dev/praw/issues/243, accessed on Jan. 24, 2018.
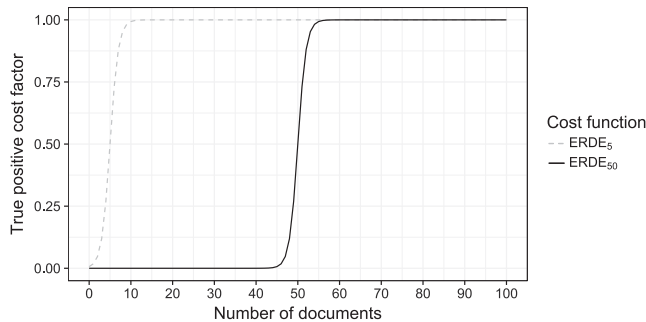7. According to the organizers, this will already be done for the eRisk 2018 task.

Fig. 1. Plot of the true positive cost factor $lc_o(k)$ for $ERDE_5$ and $ERDE_{50}$.

TABLE 2
Scores for Perfect Predictions of the eRisk 2017 Test Data after Reading n Chunks According to the Original $ERDE_o$ and the Newly Proposed $ERDE_o^\%$ Score

n	$ERDE_5$	$ERDE_{50}$	$ERDE_{20}^\%$	$ERDE_{50}^\%$
1	10.60%	3.74%	0.00%	0.00%
2	12.00%	5.37%	6.48%	0.00%
5	12.85%	8.47%	12.97%	6.48%
10	12.97%	10.48%	12.97%	12.97%

decision d submitted for a user after reading k of his messages, $ERDE_o$ is defined as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{for false positives (FP)} \\ c_{fn} & \text{for false negatives (FN)} \\ lc_o(k) \cdot c_{tp} & \text{for true positives (TP)} \\ 0 & \text{for true negatives (TN)}. \end{cases} \quad (1)$$

The values of c_{fp} and c_{fn} can be used to adjust the severity of false positives and false negatives to the given domain, while c_{tp} defines how late predictions of positive cases are punished. For the eRisk 2017 task, c_{fn} was set to 1, c_{fp} to $\frac{n_{tp}}{n_u}$, with n_{tp} denoting the number of positive cases in the test data and the total number of test users n_u. Finally, c_{tp} was set to 1 in order to treat late predictions equally to no prediction at all [54, p. 5]. The function $lc_o(k)$ determines after how many messages k the cost for true positives starts to grow and is defined as:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}}, \quad (2)$$

where the free parameter o controls around which point this logistic sigmoid function is centered. Results of the eRisk 2017 task were evaluated based on $ERDE_5$, $ERDE_{50}$, and F_1 score.

Since the results given for the baseline experiments in the original paper [53, p. 11] were obtained by using systems that could submit a prediction after reading each message per user separately, they cannot be compared to results of the actual eRisk task that required to read a whole chunk of between one and 200 messages per user. As Fig. 1 illustrates, this means that depressed users with about ten and more (for $ERDE_5$) or about 55 and more (for $ERDE_{50}$) messages per chunk basically cannot be predicted correctly because the cost would be very close to c_{fn}. Table 2 shows the $ERDE_o$ scores of perfect predictions ($F_1 = 1.0$) submitted after reading n chunks with no predictions submitted in the chunks before this one. It also includes the corresponding scores obtained from $ERDE_o^\%$ described at the end of this section. The scores obtained for $n = 1$ are the best possible $ERDE_o$ scores for this task, while $n = 10$ gives the best possible scores for a system that has read all messages. Only predicting the 18 depressed users with less than ten messages per chunk as early as possible and predicting every other user as negative, results in an F_1 score of 0.51 (1.0 precision and 0.35 recall) but still obtains an $ERDE_5$ score of 10.61 percent and $ERDE_{50}$ of 8.48 percent. The additional F_1 score is therefore especially important to evaluate systems in the general task of depression detection. To achieve better $ERDE_o$ scores, systems not only have to be optimized for this task but also need

optimized prediction thresholds to make early predictions without too many false positives. This twofold optimization makes this task especially challenging.

All experiments described in this paper are based on the exact same training and test data as the eRisk 2017 task and also process it by evaluating the same chunks of test data in chronological order. This ensures that the results are directly comparable to those of the pilot task.

Nevertheless, as the detailed look at the $ERDE_o$ function shows, there is a need to modify this function especially for future work with data in chunks. A first modification of $ERDE$ has already been proposed by Loyola et al. [67] and, in addition to making the score usable for multi-class problems, it mainly consists of an altered cost function defined now as:

$$lc_o(k) = \frac{k}{o}, \quad (3)$$

where o is no longer used to parameterize the cost but equals the number of documents per user and k is the number of documents already read for this user. This ensures that the cost is actually based on the proportion of read documents instead of a fixed number. Still, there is no way to parameterize this function and it immediately grows linearly, without any way to predict a subject correctly with a cost of zero.

We therefore propose a modification of the original sigmoid cost function:

$$lc_o(p) = 1 - \frac{1}{1 + e^{p-o}} \quad (4)$$

$$p = \frac{100 \cdot k}{n_d}, \quad (5)$$

where n_d is the total number of documents per user and k still equals to the number of documents already read per user. The cost can still be parameterized by using $o \in [0, 100]$ to make the cost grow around the point where o percent of data has been read. This results in a more intuitive cost that grows equally for all users independent of their number of messages. The newly proposed error function based on $lc_o(p)$ is denoted as $ERDE_o^\%$ and is evaluated in addition to the original function in Section 6. Table 2 shows how this score compares to the original one for perfect predictions of the eRisk 2017 test data. Since at least 10 percent of the messages per user have to be read and $o = 18$ is the minimum natural value to achieve an error of 0.00 percent, results are shown for $ERDE_{20}^\%$ instead of $ERDE_5^\%$.

Simultaneously to this work, another alternative to $ERDE$ has been proposed by a team that contributed to the eRisk 2017 task as well [68]. Their $F_{latency}$ score is based on

multiplying the standard F_1 score by a factor that is based on the latency of a system, defined as the median number of posts the system has read before predicting the positive cases. In addition, they substitute the sigmoid cost function of $ERDE_o$ by a function that increases more slowly and calculates to a penalty of 0.5 for the median number of posts in the dataset. Because this score is also tied to the absolute number of read messages, the variance of the available messages per user in the eRisk data would lead to the same problems as described above for $ERDE_o$.

4　LINGUISTIC METADATA

Augmenting the classification of text sequences with user-level metadata was one of the main ideas in this team's previous work for the eRisk task at CLEF 2017 [63]. This section builds upon this previously described set of metadata features and is aimed to further describe and extend it. All text-based metadata features are extracted from a concatenation of the *text* and *title* field (see Section 3.1) of each message, apart from obvious exceptions like the average length of these two fields. All features were calculated separately for each document of a user and then either averaged or summed up as described below.

4.1　Word and Grammar Usage

Several features based on counts of specific words or parts of speech (POS) have already been used for this team's work at the eRisk 2017 task and have been examined in the corresponding paper [63]. As described in Section 2.1, effects of depression on word and grammar usage are well-known and can include, for example, an increased usage of pronouns—especially personal pronouns—, the word "I" in particular, and verbs in past tense. Based on these previous findings, occurrences of the word "I" were counted separately in the text and title of each message in the dataset. In addition, past tense verbs, personal pronouns, and possessive pronouns were counted in the concatenation of text and title by utilizing the default POS tagger of the Python NLTK framework.[8] As an alternative to this approach, a total of 93 lexicon-based features can be obtained from the LIWC 2015 tool [20], [21]. Besides features referring to a specific POS as well, this also includes categories like emotions, informal language, or time orientations. LIWC also calculates four summary variables that represent the authenticity, emotional tone, confidence or leadership, and the amount of analytical thinking of a text. Section 4.4 describes which of these features have been utilized for the experiments of this work. For future work, this approach could likely be enhanced by utilizing more modern POS tagging approaches [69], [70] or a POS tagger that was trained specifically on a social media text domain [71]. The five final word usage features have also already been used for the participation in eRisk 2017 and aim to count very specific, hand-picked phrases that could be a strong indicator of positive cases. They count the occurrences of the exact phrases "my depression", "my anxiety", and "my therapist" as well as names of some common antidepressants[9] and

variations of the phrase "I was diagnosed with depression".[10] These very explicit features are less aimed at finding early indications of depression but at predicting the obvious cases correctly, which is important for the given task as well. In contrast to the other metadata features, this count is summed up over all documents of a user to make this a strong feature even if only present in few or a single document.

4.2　Readability

Measuring the readability or complexity of written text is a well-established idea and various different measures exist, while most of them return a result that corresponds to the school years in the US needed to understand a text. The given dataset cannot really be used as an indicator whether depressed persons write more or less complex texts because of the general difference of text quality between the classes. Since the control subjects were chosen randomly, they often differ drastically from the depressed subjects, who might use reddit to discuss their problems or generally talk to other people. Messages of the control group often simply consist of news headlines, a single short sentence, or even a single word. Readability metrics can therefore help to distinguish messages containing discussions and explanations from such simple content that is unlikely to help with the identification of depression. Several standard measures for text readability have been calculated for the text content and the four of them with the highest correlation to the class label in the training data have been selected as metadata features, namely *Gunning Fog Index* (FOG) [72], *Flesch Reading Ease* (FRE) [73], *Linsear Write Formula* (LWF)[11] [74], and *New Dale-Chall Readability* (DCR) [75], [76].

4.3　Emotions and Sentiment

As sentiment analysis is focussed on extracting opinions, affects, and emotions from written texts [33], it seems natural that knowledge from this area can also be very useful to find emotional statements in the field of mental health text classification. Especially the emotions authors express towards their personal situation could be an important indicator. While it would be possible to use the output of any state-of-the-art sentiment classification model [77] as an additional feature, this work has focussed on the use of lexicons to quickly analyze the general helpfulness of sentiment features in this dataset. First of all, the already described LIWC tool includes two features for positive and negative emotions and separate features that indicate anxiety, anger, or sadness. In addition to this, the NRC Emotion Lexicon [78] and two general sentiment lexicons, namely the Opinion Lexicon[12] and the VADER Sentiment Lexicon [79], have been used. There also exist several other lexicons that have not been evaluated, for example from the World Well-Being Project at University of Pennsylvania.[13] The NRC Emotion Lexicon contains 14,182 words that can be flagged as positive or negative and as

---

8. http://www.nltk.org/book/ch05.html, accessed on Mar. 13, 2018.

9. https://www.webmd.com/depression/guide/depression-medications-antidepressants, accessed on Mar. 14, 2018.

10. Only including the word "I" and an explicit diagnosis as in "I've been diagnosed with anxiety and depression" or "I was diagnosed with major depressive disorder."

11. Originally developed by the U.S. Air Force without any publicly available references.

12. https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon, accessed on Mar. 1, 2018.

13. http://www.wwbp.org/lexica.html, accessed on Mar. 1, 2018.

TABLE 3
User Based Metadata Features Used in Combination
with the Selected LIWC Features

Feature	Type	Description
"I" in the text	avg.	only the word "I"
"I" in the title	avg.	———— | |————
Possessive pronouns	avg.	based on POS tagging
Personal pronouns	avg.	———— | |————
Past tense verbs	avg.	———— | |————
4 readability scores	avg.	FOG, FRE, LWF, and DCR
Month of the writings	avg.	
Text length	avg.	
Title length	avg.	
Depression	sum	"my depression"
Anxiety	sum	"my anxiety"
Therapist	sum	"my therapist"
Diagnosis	sum	e.g., "I was diagnosed with depression"
Antidepressants	sum	e.g., "zoloft" and "paxil"

belonging to one or more of the emotions anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The VADER lexicon includes 7,517 terms (including emoticons) and their mean sentiment value based on the judgement of ten human annotators on a scale between −4 (extremely negative) to 4 (extremely positive). Finally, the Opinion Lexicon consists of two lists with 2,006 positive and 4,783 negative words. The corresponding counts or scores obtained from these lexicons for the eRisk 2017 dataset were again averaged over all documents of a user. Unfortunately, for this specific dataset no relevant correlation between these features and the class label could be observed. Indeed, the positive (depressed) class contains slightly more emotions and sentiments of all kinds, which might again indicate the general difference of text quality and content between the depressed subjects and the control group. As the emotion and sentiment features were of no use in this specific case, they were not included in the final set of metadata features used in the experiments of this work. Nevertheless, it can be assumed that they would be more meaningful when used with a text corpus that generally included more sentiment and emotion in both classes by using a control group that more closely resembles the target group.

4.4 Metadata Feature Summary

Table 3 summarizes the 17 metadata features that have been described above excluding the features obtained from LIWC. In addition to these, the ten LIWC features with the highest correlation to the class label in the training data have been selected. Because the LIWC lexicon includes several variations, misspellings, and abbreviations, it is accepted that some of these features already occur in the previously described feature set and therefore introduce a slight redundancy. The selected LIWC features are the number of function words, variations of the word "I" (e.g., including abbreviations containing "I" as well as "me" and "myself"), all pronouns, personal pronouns, verbs, words indicating a cognitive process, words with a focus on the present, the total number of lexicon words found, and the two calculated summary variables indicating analytical thinking and authenticity. To build the user-level metadata vector for the experiments described in Section 6, these features were again averaged over all

documents of the same user. The described metadata features result in a 27-dimensional vector per user. The concatenated feature vectors of all users are standardized as described in Section 6.1 before being used as input to a classifier. The five counts of specific terms are transformed into boolean flags by representing a value above 0 as 1 and a value of 0 as −1, similar to the previous work using LSTM networks [63]. As the experiments will show, this set of metadata features alone can lead to very good results on the eRisk 2017 dataset.

5 NEURAL NETWORK MODELS

The following sections are used to describe the neural network models that were used in the experiments of this work. All of these models are based on a document vectorization using neural word embeddings. The general concept of word embeddings and the specific models utilized in this case is described in Section 5.1. Afterwards, the following section is used to explain the type of network and the model architecture that was implemented for the experiments.

5.1 Word Embeddings

Neural word embeddings have become a popular and efficient way to model words and interactions between them for purposes like text classification tasks. They date back to the concept of distributed word representations [80] that, in contrast to local representations, do not handle each word separately with a single neuron, but use several neurons to represent a word and let each neuron be part of the description of several words. This enables distributed representations to learn general concepts of language instead of just independent word representations. One of the most important and still popular methods to train word embeddings was published by Google as word2vec [81], [82], which consists of two neural network architectures—namely the Continuous Bag of Words (CBoW) and the (Continuous) Skip-gram (SG) architecture. The concept of word2vec was developed further by Facebook and published as fastText in 2017 [83], [84], [85], which also directly offers text classification. While being based on the same two model architectures as word2vec, fastText represents words as bags of character n-grams and thus allows to obtain vectorizations even for unknown words. A different approach to learn word embeddings, GloVe, has been published by researchers of the Stanford NLP group [86]. GloVe aims to combine the advantages of local context window approaches like Skip-gram with those of global matrix factorization models like Latent Semantic Analysis (LSA).

Pre-trained word vectors obtained from large corpora are available for both fastText[14] and GloVe.[15] For this work, a 50-dimensional GloVe model trained on Wikipedia and the Gigaword 5 news corpus as well as a 300-dimensional GloVe model trained on Common Crawl were chosen. In addition, three 300-dimensional pre-trained fastText models based on similar corpora were used.

Finally, to also examine word embeddings that better fit the domain of reddit messages (or social media platforms in general), an own fastText model was trained on a dataset

---

14. https://fasttext.cc/docs/en/pretrained-vectors.html and https://fasttext.cc/docs/en/english-vectors.html, accessed on Mar. 7, 2018.
15. http://nlp.stanford.edu/projects/glove, accessed on Mar. 7, 2018.

TABLE 4
Characteristics of the GloVe and fastText Word Embeddings Used in this Paper

Model	Dimension	Corpus tokens (in billion)	Word vectors (in milllion)	Words of eRisk 2017
GloVe Wiki + News	50	6	0.4	81.9%
GloVe Crawl	300	42	2	93.2%
fastText Wiki	300	?	2.5	81.7%
fastText Wiki + News	300	16	1	79.5%
fastText Crawl	300	600	2	88.5%
fastText reddit	300	49.9	6	99.7%

Fig. 2. Nearest neighbors for six hand-picked tokens in the fastText model trained on reddit comments according to cosine similarity.

containing all reddit comments between October 2007 and May 2015.[16] The total dataset consists of about 1.7 billion messages that we preprocessed and tokenized in a way that preserves emoticons, punctuation, and words that include special characters (e.g., censored words). The preprocessing step also included to replace any references to reddit users (in the form of /u/ < username >) 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. The resulting tokens of each message were lowercased (with the exception of emoticons) and separated with a space to enable fastText to properly treat punctuation.[17] Since the dataset also contains messages written in different languages than English and a sophisticated language detection classifier would have required too much time for so many documents, a simple language detection based on stopword counts was utilized: Only messages with more English stopwords than ones from other languages[18] were retained (thus also discarding messages without any stopwords). This resulted in a final training corpus of 1.37 billion messages and a total of 49.9 billion tokens. The C++ implementation of fastText was used to train a 300-dimensional CBoW model without subword information and default values for all other parameters, which contains the 6 million unique tokens that occur at least five times in this corpus. Training took about 26 hours using 12 threads on an Intel Xeon E5-2687W 3.10 GHz and needed about 17 GB of RAM.

The characteristics of all utilized word embeddings are shown in Table 4. It also contains the amount of words each model includes of the 85,558 words that occur in writings of at least two users of the eRisk 2017 dataset after the same preprocessing and tokenization steps as described above. This set of words is used in the experiments described in Section 6.

As a qualitative analysis of the self-trained fastText model, it is possible to examine the nearest neighbors of some hand-picked exemplary tokens according to cosine similarity. Fig. 2 displays six word clouds with the corresponding example token in the middle and its ten (20 in the case of emoticons) nearest neighbors around it. These examples especially illustrate that the model trained on reddit messages is indeed able to identify similar emoticons and subreddits, which is both not possible using the pre-trained fastText or GloVe models. The closest neighbors of the depression subreddit also include terms like "suicidewatch" and "mmfb" (make me feel better), illustrating relations that were learnt to terms besides other subreddits. It has also learnt similar embeddings for antidepressants like Zoloft. While this as well as similar words to the terms "depression" and "suicidal" can also be observed using the pre-trained models, their nearest neighbors seem slightly more medical and from a more neutral perspective (e.g., "ssri" or "sertraline" close to "zoloft", "melancholia" or "insomnia" for "depression", and "deranged" or "delusional" for "suicidal") than those of the model trained on reddit. Also, especially the fastText Crawl model returns neighboring terms like "depression.This" or "depression.What", which might indicate a preprocessing problem concerning punctuation.

To further investigate the extent to which the self-trained fastText reddit model has learnt a general model of the English language, the standard word analogy dataset [81] can be used as one indicator. Table 5 compares results on the word analogy dataset published for state-of-the-art models to the results of the new model trained on reddit messages. The dataset was originally created to evaluate the word embeddings of word2vec and consists of 8.869 semantic as well as 10.675 syntactic analogy questions: Given an analogy (e.g., Athens and Greece) and a third word (e.g., Oslo), word

16. https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/, accessed on Mar. 7, 2018.
17. https://github.com/facebookresearch/fastText/issues/333, accessed on Mar. 7, 2018.
18. based on the 2,400 stopwords for 11 languages of the NLTK Python package, http://www.nltk.org/book/ch02.html#annotated-text-corpora, accessed on Mar. 7, 2018.

TABLE 5
Semantic, Syntactic, and Total Accuracies on the Word
Analogy Dataset for 300-Dimensional Word Embeddings

Model	Sem.	Syn.	Total
word2vec [82]	63	58	61
GloVe Crawl [86]	82	69	75
fastText Wiki [84]	78	75	76
fastText Wiki + News [85]	**90**	**84**	**87**
fastText Crawl [85]	87	82	85
fastText reddit	56	74	66

embeddings have to return the fourth word (Norway in this case) as closest vector to the result of $vec("Greece") - vec("Athens") + vec("Oslo")$ according to cosine similarity. While the results are far from the ones obtained by the state-of-the-art fastText models trained on Wikipedia and news articles, especially the result in the syntactic category illustrates that even the training on these much less formal documents has lead to a decent model of the English language.

5.2 Neural Network Architecture

Convolutional Neural Networks (CNN) [87] have been utilized to achieve outstanding results especially in the area of image classification and are generally viable for data with a grid-like structure [88]. Recently, studies have shown that they can also be used effectively for text classification tasks [89]. Fig. 3 displays the architecture of the simple CNN used for the experiments described in this paper, which is based on the one-layer CNN for sentence classification described by Zhang and Wallace [89]. Similar to this sentence classification network, it consists of only a single CNN layer but uses a total of 100 filters with a height of 2 and a width corresponding to the word vector dimensions. Concatenated Rectified Linear Units (CReLU) [90] are used as activation function for the convolutional layer as well as for the fully-connected layers, resulting in twice as many outputs due to the concatenation with the negated activation. 1-max pooling is used to obtain a single scalar from each filter, which results in a 200-dimensional vector due to the CReLU activation. This output is then propagated through three fully-connected layers with, again, CReLU activation, of which the first one applies dropout to its output. The fourth and final layer

applies softmax to the output. As input, the network receives each document of a user individually in the form of the 100 first word vectors per document, while using zero-padding for documents with less than 100 words. This results in a 100×50 or 100×300 dimensional input matrix depending on the used word embedding as described in the previous section. The limitation to 100 words (or even less) is possible as the number of words per document 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. As this results in a separate output for each document per user, the 98th percentile of these outputs is used as final prediction for the user. This value is chosen instead of the mean to give more weight to documents with a higher probability.

In addition to the described model architecture, it has been tried to directly use the metadata features as a second input to the network. An approach similar the the previous one at eRisk 2017 [63], where the metadata features were fed through a fully-connected layer and then concatenated with the output of an LSTM layer, did not lead to better results than just using the text input. The same applies to the idea to use the final 1×50 dimensional vector before the softmax layer as additional input for a metadata classifier like a logistic regression. Since the results of networks using the metadata differed only marginally from the text-only network described above, only results of the latter will be reported in the following to reduce the complexity. Future work will be needed to explore better ways to directly merge the CNN output with the metadata. This could, for example, be done by implementing a dedicated fusion component into the neural network, similar to work done for gender identification based on texts and images at the CLEF 2018 PAN workshop [91]. In this work, a simple late fusion ensemble will show that the best results so far can indeed be achieved by combining these features.

6 EXPERIMENTS

This section is used to describe the experiments done based on the convolutional neural network and the metadata features as well as their results. Results are compared to the best published results during the eRisk 2017 task as well as other results obtained after the ground truth was released.
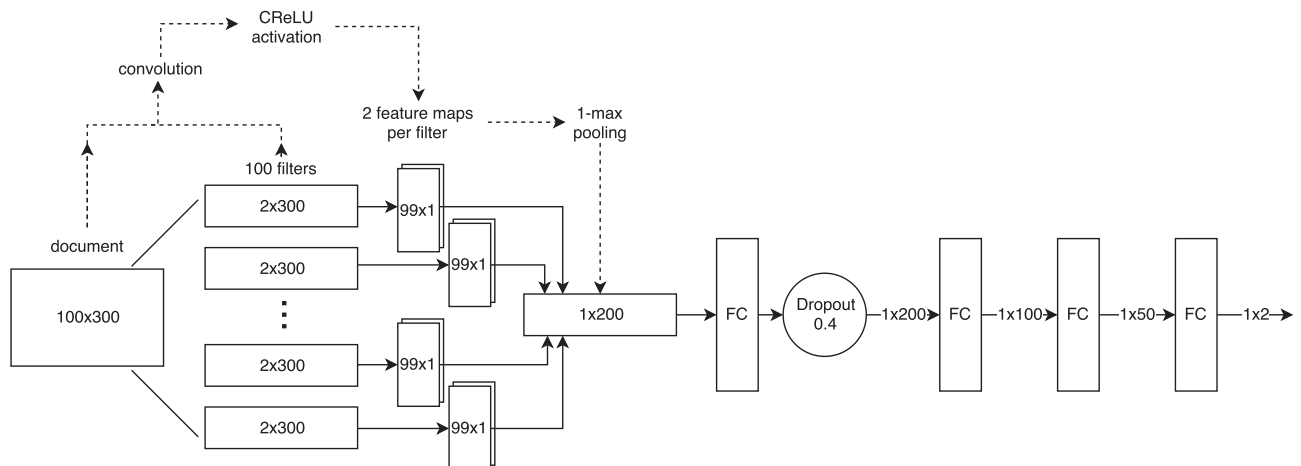

Fig. 3. Network architecture used for the convolutional neural network models described in this paper. Adapted from [89].

TABLE 6
$ERDE_o$ Scores, F_1 Score, Precision, and Recall of the
Utilized Models and Previous Publications

Model	$p >$	$ERDE_5$	$ERDE_{50}$	F_1	P	R
UNSLA [54]		13.66	9.68	0.59	0.48	0.79
FHDO-BCSGA [54]		12.82	9.69	0.64	0.61	0.64
FHDO-BCSGB [54]		12.70	10.39	0.55	0.69	0.46
TVT-NB [62]		13.13	8.17	0.54	0.42	0.73
TVT-RF [62]		12.30	8.95	0.56	0.54	0.58
GloVe W+N	0.5	12.95	7.57	0.63	0.56	0.73
GloVe Crawl	0.7	12.98	8.59	0.63	0.58	0.69
fastText Wiki	0.6	13.06	8.17	0.57	0.47	0.71
fastText W+N	0.55	13.11	7.95	0.60	0.49	0.77
fastText Crawl	0.6	13.01	8.60	0.64	0.60	0.67
fastText reddit	0.7	13.52	8.04	0.62	0.51	0.79
fastText reddit	0.8	12.71	9.23	0.56	0.63	0.50
Meta LR	0.35	12.65	8.57	0.66	0.59	0.73
Meta LR	0.55	12.35	9.86	0.65	0.72	0.60
Meta LR Wait	0.35	13.32	11.33	**0.73**	**0.77**	0.69
G W+N + Meta LR	0.45	12.34	8.93	0.71	0.72	0.69
fT Wiki + Meta LR	0.35	13.52	**7.29**	0.55	0.41	**0.85**
fT Wiki + Meta LR	0.5	**12.13**	8.77	0.71	0.71	0.71
fT reddit + Meta LR	0.55	12.46	8.77	0.67	0.69	0.65

The second column displays the prediction threshold.

The scores of each model are reported according to $ERDE_5$, $ERDE_{50}$, and F_1, which are the official scores of this task, and also based on the newly proposed $ERDE_{20}^\%$, $ERDE_{50}^\%$, and $F_{latency}$.

6.1 Experiment Setup

For the experiments conducted during this work, the same process used during the eRisk 2017 task was reproduced: The available test documents were processed in the same ten chunks that contain 10 percent of the writings obtained from each user. Training is done once on the full training dataset. Afterwards, test chunks are processed in sequential order, while the documents of the previous chunks are always used again. The only exception to this process is the model called "Meta LR Wait" in the following evaluation section, which is a logistic regression based on metadata features that was configured to only submit a prediction after the final chunk. Similar "waiting" models were also utilized by some teams during eRisk 2017 and can be interesting to evaluate the possible F_1 score, while neglecting the early detection aspect and therefore the $ERDE_o$ scores. Since the models based on metadata use features averaged over all documents of the same user, they were also calculated for each chunk separately, again using documents from earlier chunks as well. An additional parameter for the early detection models is the prediction threshold that determines whether a model is confident enough to predict a subject as positive (depressed) or whether it waits for more data. While these thresholds were based on cross-validation using the training data for this team's participation in the eRisk 2017 task and included the number of documents already read in multiple threshold levels [63], the experiments in this work are based on a single threshold value that achieved the best test result for the specific model. This is likely to lead to an overfitting on the specific test data but also allows to compare the best possible results of the utilized models. Generally, prediction thresholds between 0.5 and 0.7

lead to a balanced result in all scores, while higher thresholds can often maximize $ERDE_5$ but severely decrease F_1. This fits the observations described in Section 3.2: Since the correct prediction of so few depressed test subjects actually has an effect on $ERDE_5$, it is often best to submit fewer predictions overall and therefore simply minimize false positives. Negative (non-depressed) predictions were only submitted after seeing the final chunk. The models based on user metadata features all utilized the same logistic regression classifier. The 27 features described in Section 4.4 were first standardized to have a mean of 0 and unit variance, with exception of the boolean flag features that already have a value of either -1 or 1. The resulting scaled feature vector was then used to train a logistic regression classifier and later predict probabilities for the test subjects of each chunk.

6.2 Evaluation

Table 6 displays the results achieved in this work in comparison to previously published results for the same dataset and task. The first three rows in this table represent the best results during the eRisk 2017 task and are therefore solely optimized based on cross-validation over the training data, while the next two results have been achieved after the ground truth was published. All results after these have been achieved as part of this work. The models corresponding to the name of a word embedding refer to a CNN using this embedding as input vectorization, the models named "Meta LR" refer to the logistic regression based on metadata, and the final four results were obtained by calculating the mean of the metadata probabilities and the neural network output. Although these outputs have not been calibrated (e.g., by using Platt scaling [92]), this simple late fusion ensemble lead to the best achieved $ERDE_o$ scores and recall. As expected, the best overall F_1 score could be obtained by waiting for the last chunk and only then submitting predictions based on the metadata LR. Interestingly, this model would still have achieved the seventh best $ERDE_{50}$ score in the eRisk 2017 task out of 30 submissions, which again illustrates how difficult $ERDE_o$ is to interpret because it is based on the absolute number of documents. The prediction thresholds have been chosen to represent the best possible $ERDE_o$ scores that still include a viable F_1 score. A second threshold has been reported for the self-trained fastText reddit model and the metadata LR to illustrate to which extent slightly different thresholds can have an effect on $ERDE_o$ scores. As already described, especially optimizing $ERDE_5$ often includes impairing F_1 score. The reported results contain the best scores published for this task so far and, importantly, achieve a balanced result among all scores. Although $ERDE_o$ and F_1 are difficult to maximize simultaneously, the results show that the described models are able to do so.

In addition to comparing the achieved results to previously published results for this task, Table 7 shows how the same models with the same prediction thresholds would have scored according to the newly proposed $ERDE_o^\%$ score as well as the $F_{latency}$ score from [68]. While the "Meta LR Wait" model is now scored equally bad in both criteria because it had to read all documents, the CNN scores now tend to be better than the ones obtained for the metadata models alone. Still, the best overall $ERDE_o^\%$ scores could be

TABLE 7
Scores of the above Models According to the Proposed
$ERDE_o^\%$ Criterium and the $F_{latency}$ Measure from [68]

Model	$p >$	$ERDE_{20}^\%$	$ERDE_{50}^\%$	$F_{latency}$
GloVe W+N	0.5	8.70	7.08	**0.52**
GloVe Crawl	0.7	9.44	6.58	0.44
fastText Wiki	0.6	8.18	6.69	0.46
fastText W+N	0.55	8.34	6.10	0.49
fastText Crawl	0.6	9.60	7.23	0.48
fastText reddit	0.7	8.53	5.53	0.51
fastText reddit	0.8	10.21	9.21	0.35
Meta LR	0.35	9.32	7.32	0.43
Meta LR	0.55	10.74	8.74	0.40
Meta LR Wait	0.35	13.32	13.32	0.26
G W+N + Meta LR	0.45	9.68	7.56	0.45
fT Wiki + Meta LR	0.35	**6.40**	**4.78**	0.49
fT Wiki + Meta LR	0.5	9.21	7.47	0.48
fT reddit + Meta LR	0.55	9.46	7.47	0.45

achieved by the same ensemble. The additional models with higher thresholds that were previously included to obtain a better $ERDE_5$ score (namely fastText reddit with $p > 0.8$ and Meta LR with $p > 0.55$) now result in the worst overall $ERDE_o^\%$ scores next to the waiting model. This again indicates that especially optimizations of $ERDE_5$ do not necessarily mean a better classification result. The 50-dimensional GloVe model achieves the best $F_{latency}$ score, which is also better than the best score reported in the original paper (0.389) for the same dataset [68].

7 CONCLUSION

This work has been used to examine the currently popular $ERDE_o$ metric for early detection tasks in detail and has shown that especially $ERDE_5$ is not a meaningful metric for the described shared task. Only the correct prediction of few positive samples has an effect on this score and the best results can therefore often be obtained by only minimizing false positives. A modification of this metric, namely $ERDE_o^\%$, has been proposed that is better interpretable in the case of shared tasks that require information to be read in chunks. Exemplary scores using this score have been shown in comparison to $ERDE_o$ scores for the experiments in this work.

Previous experiments for the eRisk 2017 task for early detection of depression have been continued by examining additional user-level metadata features and evaluating a convolutional neural network as text-based depression classifier. State-of-the-art results have been reported for the eRisk 2017 dataset using these two approaches. A new fastText word embedding has been trained on a large corpus of reddit comments. The analysis of the resulting word vectors has shown that the model has learnt some features specific to this domain and is viable for general syntactic questions in the English language as shown based on the standard word analogy task.

As the results presented in this paper are optimized to obtain the best performance on the eRisk 2017 task for comparison to previously published results and among these models, future work will have to show how these models perform on yet unseen data. This has first been done during the eRisk 2018 task [93], which used the old dataset as training data and contained 820 new test subjects. In addition, eRisk

2018 contained an additional task aimed at the early detection of anorexia that this team has also participated in. The five submitted predictions achieved the best F_1 and $ERDE_{50}$ scores in both tasks and the CNN without metadata in particular achieved the best results in the new anorexia task [94]. The same working notes paper for this second participation has also been used to evaluate the modified $ERDE_o^\%$ metric for all participants and again shows how especially the original $ERDE_5$ metric favors systems that correctly predict test users with only few documents in total regardless of their overall performance.

As the detailed look at the current $ERDE_o$ metric has shown, one priority of future work in this area should be to agree on a new metric for early detection tasks like eRisk. Ethical issues in this area of research have been reviewed and should find more attention as well. Possibilities to publish the fastText model trained on reddit comments still have to be examined. Concerning the models presented in this work, additional experiments will be necessary to find better ways to integrate the metadata features directly into the neural network. On the other hand, utilizing ensembles of more than just two models and calibrating the resulting probabilities seems promising. Combining word embeddings of two models in a single neural network has also not been evaluated yet. Another possible improvement would be to use recently published language modeling methods like BERT as input for the network and to compare a self-trained model using this approach to the fastText word embeddings of this work.

REFERENCES

[1] *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization, 2017, https://www.who.int/mental_health/management/depression/prevalence_global_health_estimates/en/

[2] A. T. Beck and B. A. Alford, *Depression: Causes and Treatment. 2nd ed.* Philadelphia, PA, USA: Univ. Pennsylvania Press, 2009.

[3] *Key Substance Use and Mental Health Indicators in the United States: Results from the 2016 National Survey on Drug Use and Health*, Rockville, MD: Center for Behavioral Health Statistics and Quality: Substance Abuse and Mental Health Services Administration, 2017. [Online]. Available: https://www.samhsa.gov/data/

[4] E. S. Paykel, "Basic concepts of depression," *Dialogues Clinical Neuroscience*, vol. 10, no. 3, pp. 279–289, 2008.

[5] *Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2015*. World Health Organization, 2016, https://www.who.int/entity/healthinfo/global_burden_disease/GHE2015_Deaths_Global_2000_2015.xls

[6] J. Alonso, M. Codony, V. Kovess, M. C. Angermeyer, S. J. Katz, J. M. Haro, G. De Girolamo, R. De Graaf, K. Demyttenaere, G. Vilagut, et al., "Population level of unmet need for mental healthcare in europe," *British J. Psychiatry*, vol. 190, no. 4, pp. 299–306, 2007.

[7] P. S. Wang, M. Angermeyer, G. Borges, R. Bruffaerts, W. T. Chiu, G. De Girolamo, J. Fayyad, O. Gureje, J. M. Haro, Y. Huang, et al., "Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative," *World Psychiatry*, vol. 6, no. 3, 2007, Art. no. 177.

[8] A. Rahman, S. U. Hamdani, N. R. Awan, R. A. Bryant, K. S. Dawson, M. F. Khan, M. M.-U.-H. Azeemi, P. Akhtar, H. Nazir, A. Chiumento, et al., "Effect of a multicomponent behavioral intervention in adults impaired by psychological distress in a conflict-affected area of pakistan: A randomized clinical trial," *JAMA*, vol. 316, no. 24, pp. 2609–2617, 2016.

[9] G. Schomerus, H. Matschinger, and M. C. Angermeyer, "The stigma of psychiatric treatment and help-seeking intentions for depression," *Eur. Archives Psychiatry Clinical Neurosci.*, vol. 259, no. 5, pp. 298–306, 2009.

[10] R. Whitley and R. D. Campbell, "Stigma, agency and recovery amongst people with severe mental illness," *Social Sci. Med.*, vol. 107, pp. 1–8, 2014.

[11] K. Gowen, M. Deschaine, D. Gruttadara, and D. Markey, "Young adults with mental health conditions and social networking websites: Seeking tools to build community," *Psychiatric Rehabil. J.*, vol. 35, no. 3, pp. 245–250, 2012.

[12] J. A. Naslund, S. W. Grande, K. A. Aschbrenner, and G. Elwyn, "Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using youtube," *PLOS One*, vol. 9, no. 10, 2014, Art. no. e110171.

[13] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels, "The future of mental health care: Peer-to-peer support and social media," *Epidemiology Psychiatric Sci.*, vol. 25, no. 2, pp. 113–122, 2016.

[14] M. Berger, T. H. Wagner, and L. C. Baker, "Internet use and stigmatized illness," *Social Sci. Med.*, vol. 61, no. 8, pp. 1821–1827, 2005.

[15] W. Bucci and N. Freedman, "The language of depression," *Bulletin Menninger Clinic*, vol. 45, no. 4, pp. 334–358, 1981.

[16] W. Weintraub, *Verbal Behavior: Adaptation and Psychopathology*. New York, NY, USA: Springer Publishing Company, 1981.

[17] S. Rude, E.-M. Gortner, and J. Pennebaker, "Language use of depressed and depression-vulnerable college students," *Cognition Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[18] D. Smirnova, E. Sloeva, N. Kuvshinova, A. Krasnov, D. Romanov, and G. Nosachev, "Language changes as an important psycho-pathological phenomenon of mild depression," presented at the 21st Eur. Congr. Psychiatry, Nice, France, 2013.

[19] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Sci.*, vol. 6, no. 4, pp. 529–542, 2018.

[20] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[21] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annu. Rev. Psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[22] H. A. Schwartz, S. Giorgi, M. Sap, P. Crutchley, L. Ungar, and J. Eichstaedt, "DLATK: Differential language analysis toolkit," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2017, pp. 55–60. [Online]. Available: http://aclweb.org/anthology/D17-2010

[23] D. Hovy and S. L. Spruit, "The social impact of natural language processing," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, vol. 2, pp. 591–598.

[24] J. L. Leidner and V. Plachouras, "Ethical by design: Ethics best practices for natural language processing," in *Proc. 1st ACL Workshop Ethics Natural Lang. Process.*, 2017, pp. 8–18.

[25] A. Benton, G. Coppersmith, and M. Dredze, "Ethical research protocols for social media health research," in *Proc. 1st ACL Workshop Ethics Natural Lang. Process.*, 2017, pp. 94–102.

[26] C. P. Escartín, W. Reijers, T. Lynn, J. Moorkens, A. Way, and C.-H. Liu, "Ethical considerations in NLP shared tasks," in *Proc. 1st ACL Workshop Ethics Natural Lang. Process.*, 2017, pp. 66–73.

[27] J. Mikal, S. Hurst, and M. Conway, "Ethical issues in using Twitter for population-level depression monitoring: A qualitative study," *BMC Med. Ethics*, vol. 17, Apr. 2016, Art. no. 22.

[28] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[29] M. E. Maron, "Automatic indexing: An experimental inquiry," *J. Assoc. Comput. Machinery*, vol. 8, no. 3, pp. 404–417, 1961.

[30] P. J. Hayes, P. M. Andersen, I. B. Nirenburg, and L. M. Schmandt, "TCS: A shell for content-based text categorization," in *Proc. 6th IEEE Conf. Artif. Intell. Appl.*, 1990, pp. 320–326.

[31] T. M. Mitchell, et al., *Machine Learning*. Boston, MA, USA: McGraw-Hill, 1997.

[32] F. Sebastiani, "Machine learning in automated text catego-rization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[33] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[34] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.

[35] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 417–424.

[36] R. Johnson and T. Zhang, "Deep pyramid convolutional neural networks for text categorization," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2017, vol. 1, pp. 562–570.

[37] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2018, vol. 1, pp. 2227–2237.

[38] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Vol. 1: Long Papers)*, 2018, vol. 1, pp. 328–339.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv: 1810.04805, 2018.

[40] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing Twitter for public health," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, vol. 20, pp. 265–272.

[41] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proc. 5th Annu. ACM Web Sci. Conf.*, 2013, pp. 47–56.

[42] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, vol. 13, pp. 1–10.

[43] M. Nadeem, M. Horn, G. Coppersmith, and S. Sen, "Identifying depression on Twitter," arXiv:1607.07384, 2016.

[44] Y.-H. Huang, L.-H. Wei, and Y.-S. Chen, "Detection of the prodromal phase of bipolar disorder from psychological and phonological aspects in social media," arXiv: 1712.09183, 2017.

[45] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proc. 2nd Workshop Comput. Linguistics Clinical Psychology: From Linguistic Signal Clinical Reality*, 2015, pp. 31–39.

[46] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion Behavioral Sci.*, vol. 18, pp. 43–49, 2017.

[47] J. Á. Bitsch, R. Ramos, T. Ix, P. G. Ferrer-Cheng, and K. Wehrle, "Psychologist in a pocket: towards depression screening on mobile phones," *Studies Health Technol. Inform.*, 2015, vol. 211, Art. no. 153–159. [Online]. Available: http://europepmc.org/abstract/MED/25980862

[48] P. G. F. Cheng, R. M. Ramos, J. Á. Bitsch, S. M. Jonas, T. Ix, P. L. Q. See, and K. Wehrle, "Psychologist in a pocket: lexicon development and content validation of a mobile-based app for depression screening," *JMIR mHealth uHealth*, vol. 4, no. 3, 2016, Art. no. e88.

[49] G. Dulac-Arnold, L. Denoyer, and P. Gallinari, "Text classification: A sequential reading approach," in *Proc. 33rd Eur. Conf. Inf. Retrieval: Advances Inf. Retrieval*, 2011, pp. 411–423.

[50] H. J. Escalante, M. Montes-y-Gómez, L. Villaseñor-Pineda, and M. L. Errecalde, "Early text classification: a naïve solution," in *Proc. 7th Workshop Comput. Approaches Subjectivity Sentiment Social Media Anal.*, 2016, pp. 91–99.

[51] M. Torii, L. Yin, T. Nguyen, C. T. Mazumdar, H. Liu, D. M. Hartley, and N. P. Nelson, "An exploratory study of a text classification framework for internet-based surveillance of emerging epidemics," *Int. J. Med. Inform.*, vol. 80, no. 1, pp. 56–66, 2011.

[52] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 1395–1405. [Online]. Available: https://doi.org/10.1145/2736277.2741637

[53] D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *Proc. 7th Int. Conf. CLEF Assoc.*, 2016, pp. 28–39.

[54] D. E. Losada, F. Crestani, and J. Parapar, "CLEF 2017 eRisk overview: Early risk prediction on the internet: Experimental foundations," in *Proc. Conf. Labs Eval. Forum CLEF*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/invited_paper_5.pdf

[55] D. E. Losada, F. Crestani, and J. Parapar, "eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer International Publishing, 2017, pp. 346–360.

[56] H. Almeida, A. Briand, and M.-J. Meurs, "Detecting early risk of depression from social media user-generated content," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_127.pdf

[57] A. A. Farias-Anzaldua, M. Montes-Y-Gómez, A. P. Lopez-Monroy, and L. C. Gonzalez-Gurrola, "UACH-INAOE participation at eRisk2017," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_136.pdf

[58] I. A. Malam, M. Arziki, M. N. Bellazrak, F. Benamara, A. El Kaidi, B. Es-Saghir, Z. He, M. Housni, V. Moriceau, J. Mothe, and F. Ramiandrisoa, "IRIT at e-Risk," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_135.pdf

[59] F. Sadeque, D. Xu, and S. Bethard, "UArizona at the CLEF eRisk 2017 pilot task: Linear and recurrent models for early depression detection," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_58.pdf

[60] E. Villatoro-Tello, G. Ramírez-De-La-Rosa, and H. J. Salazar, "UAMś participation at CLEF eRisk 2017 task: Towards modelling depressed blogers," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_113.pdf

[61] M. L. Errecalde, M. L. Villegas, D. G. Funez, M. J. G. Ucelay, and L. C. Cagnina, "Temporal variation of terms as concept space for early risk prediction," in *Proc. Conf. Labs Eval Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_103.pdf

[62] M. P. Villegas, D. G. Funez, M. J. G. Ucelay, L. C. Cagnina, and M. L. Errecalde, "LIDIC - UNSL's participation at eRisk 2017: Pilot task on early detection of depression," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_107.pdf

[63] M. Trotzek, S. Koitka, and C. M. Friedrich, "Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression," in *Proc. Conf. Labs Eval. Forum*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/paper_54.pdf

[64] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, "CLPsych 2016 shared task: Triaging content in online peer-support forums," in *Proc. 3rd Workshop Comput. Lingustics Clinical Psychology*, 2016, pp. 118–127.

[65] D. N. Milne, "Triaging content in online peer-support: an overview of the 2017 CLPsych shared task," 2017. [Online]. Available: http://clpsych.org/shared-task-2017

[66] V. Lynn, A. Goodman, K. Niederhoffer, K. Loveys, P. Resnik, and H. A. Schwartz, "CLPsych 2018 shared task: Predicting current and future psychological health from childhood essays," in *Proc. 5th Workshop Comput. Linguistics Clinical Psychology: From Keyboard Clinic*, 2018, pp. 37–46. [Online]. Available: http://aclweb.org/anthology/W18-0604

[67] J. M. Loyola, M. L. Errecalde, H. J. Escalante, and M. Montes y Gomez, "Learning when to classify for early text classification," presented at the XXIII Congreso Argentino de Ciencias de la Computación, La Plata, Argentina, 2017.

[68] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 495–503. [Online]. Available: http://doi.acm.org/10.1145/3159652.3159725

[69] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv:1508.01991, 2015.

[70] J. D. Choi, "Dynamic feature induction: The last gist to the state-of-the-art," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, 2016, pp. 271–281.

[71] M. Neunerdt, B. Trevisan, M. Reyer, and R. Mathar, "Part-of-speech tagging for social media texts," *Language Processing and Knowledge in the Web*. New York, NY, USA: Springer, 2013, pp. 139–150.

[72] R. Gunning, *The Technique of Clear Writing*. New York, NY, USA: McGraw-Hill, 1952.

[73] R. Flesch, "A new readability yardstick," *J. Appl. Psychology*, vol. 32, no. 3, pp. 221–233, 1948.

[74] G. J. Christensen, "Readability helps the level," 2006. [Online]. Available: http://www.csun.edu/~vcecn006/read1.html

[75] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educational Res. Bulletin*, vol. 27, pp. 37–54, 1948.

[76] J. S. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Northampton, MA, USA: Brookline Books, 1995.

[77] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Rev.: Data Mining Knowl. Discovery*, 2018, Art. no. e1253.

[78] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[79] C. H. E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs Social Media*, 2014.

[80] G. E. Hinton, J. L. McClelland, D. E. Rumelhart, et al., "Distributed representations," *Parallel Distrib. Process.: Explorations Microstructure Cognition*, vol. 1, no. 3, pp. 77–109, 1986.

[81] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Representations Workshops Track*, 2013.

[82] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[83] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, vol. 2, pp. 427–431.

[84] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

[85] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proc. 11th Int. Conf. Language Resources Evaluation (LREC 2018)*, May 2018.

[86] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[87] Y. LeCun, "Generalization and network design strategies," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-89–4, 1989.

[88] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[89] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (Vol. 1: Long Papers)*, 2017, pp. 253–263.

[90] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, vol. 48, pp. 2217–2225.

[91] T. Takahashi, T. Tahara, K. Nagatani, Y. Miura, T. Taniguchi, and T. Ohkuma, "Text and image synergy with feature cross technique for gender identification: Notebook for PAN at CLEF 2018, " in *Proc. Conf. Labs Eval. Forum*, 2018. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_83.pdf

[92] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[93] D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, Eds. Cham, Switzerland: Springer International Publishing, 2018, pp. 343–361.

[94] M. Trotzek, S. Koitka, and C. M. Friedrich, "Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia," in *Proc. Conf. Labs Eval. Forum*, 2018. [Online]. Available: http://ceur-ws.org/Vol-2125/paper_68.pdf

Marcel Trotzek received the MSc degree in computer science from the University of Applied Sciences and Arts Dortmund, Germany, in 2018. His research interest includes area of machine learning with a focus on text classification.

Sven Koitka received the MSc degree in computer science from the University of Applied Sciences and Arts Dortmund, Germany, in 2016 and is currently working toward the PhD degree. His research interests include machine learning, in particular neural networks with a focus on computer vision.

Christoph M. Friedrich received the diploma degree in computer science from the University of Dortmund, Germany, in 1996 and the PhD degree in life science engineering from the University of Witten/Herdecke, Germany, in 2006. He joined the Computer Science Department, University of Applied Sciences and Arts Dortmund, Germany, as a professor of biomedical computer science, in 2010. From 2005 to 2010, he led the Data Mining Group in the Bioinformatics Department, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI). Besides his academic career, he has several years of industrial experience. He is author or co-author of more than 80 publications in international journals or conference proceedings. His main field of interest is in machine learning, text mining, biomedical applications of computer vision, and bioinformatics. He is a member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.