Detecting Early Onset of Depression from Social Media Text using Learned Confidence Scores

Ana-Maria Bucur
University of Bucharest, Romania
ana-maria.bucur@drd.unibuc.ro

Liviu P. Dinu
University of Bucharest, Romania
ldinu@fmi.unibuc.ro

Abstract

English. Computational research on mental health disorders from written texts covers an interdisciplinary area between natural language processing and psychology. A crucial aspect of this problem is prevention and early diagnosis, as suicide resulted from depression being the second leading cause of death for young adults. In this work, we focus on methods for detecting the early onset of depression from social media texts, in particular from Reddit. To that end, we explore the eRisk 2018 dataset and achieve good results with regard to the state of the art by leveraging topic analysis and learned confidence scores to guide the decision process. [1]

1 Introduction

Mental illnesses are a common problem of our modern world. More than one in ten people was living with mental health disorders in 2017 (Ritchie and Roser, 2018), with women being the most affected. These disorders affect people's way of thinking, mood, emotions, behaviour and their relationships with others. Most mental illnesses remain undiagnosed because of the social stigma around them.

Depression is one of the main causes of disability globally [2], it affects people of all ages. Prevention is used to reduce depression and to save the lives of people at risk of suicide, but prevention is only limited to raising awareness and programs to cultivate positive thinking in case of depression and monitoring people who attempted suicide or self-harm.

With the rise in social media use, more computational efforts are made to detect mental illnesses such as depression (De Choudhury et al., 2013) and PTSD (Coppersmith et al., 2015), but also to detect misogyny (Anzovino et al., 2018), irony and sarcasm (Khokhlova et al., 2016) from users' texts.

People tend to talk more about their emotions and mental health problems online and to seek support. The sources of mental health cues used for detection are Twitter, Facebook, Reddit and forums (Calvo et al., 2017). Reddit[3] is a social media site very similar to forums. It is organized in subreddits with specific topics, some dedicated to mental health problems. The use of throwaway accounts to maintain anonymity promotes disclosure, and users are more likely to share problems they have not discussed with anyone before. The use of these accounts makes it difficult for users to receive more social support because the majority of them are used only for one post (Calvo et al., 2017).

In this work, we choose to tackle the problem of detecting early onset of depression from users' posts on social media, specifically from Reddit. As such, we explore the eRisk 2018 dataset through topic analysis by means of Latent Semantic Indexing (Deerwester et al., 1990) and learned out-of-distribution confidence scores (DeVries and Taylor, 2018). Due to the nature of the dataset, we repurpose the learned confidence score to make a decision on whether to label the user as depressed or non-depressed or to wait for more data, as test chunks were progressively released every week.

2 Related Work

Recent studies for depression detection from text are reviewed by Guntuku et al. (Guntuku et al., 2017). People diagnosed with mental illnesses from the datasets are identified using screening surveys, self-reported posts about diagnosis from social media or by their membership in different forums related to mental health. The most used fea-

[2] https://www.who.int

[3] https://www.reddit.com

tures are topic modelling, n-grams, Linguistic Inquiry and Word Count (LIWC), emotion and metadata. The most used methods are Support Vector Machines (SVM), Logistic Regression, Random Forests and Neural Networks.

Coppersmith et al. (2016) show the differences in emoticons use between suicidal users and controls, neurotypicals using emojis with a much higher probability than a user before an attempt. Prior to the suicide attempt, the users at risk tend to use a more self-focused language, same as the people diagnosed with depression. The authors highlight different changes in post emotions before and after the suicide attempt. Users are also more likely to talk about suicide after an attempt than before it.

Sekulić et al. (2018) indicate that users diagnosed with bipolar disorders use more first-person singular pronouns, same as depressed people. They also use more words associated with emotions; words associated with positive emotions as well as words associated with negative emotions explained by alternating episodes of mania and depression.

Nalabandian el al. (2019) show that depressed persons tend to use more negative words and a self-focused language when writing about their interactions with a close romantic partner than when writing about other people around them. This is because people experience different symptoms of mental illness based on their interactions with other people.

Loveys et al. (Loveys et al., 2018) show the differences in language use of users with depression from different cultures to avoid cultural biases. Even if depression affects people all over the world, the way they experience and express it is shaped by their cultural context. Users from some ethnic groups does not address mental health issues online as much as the others and this can make the depression task more difficult. After topic modeling, the authors show that the words from each topic vary for each ethnic group, people discussing different themes relevant to their culture.

For diagnosis before the onset of the mental health disorders, Eichstaedt et al. (2018) use users' posts from Facebook to predict a future depression diagnosis. De Choudhury et al. (2013) use a classifier to predict users' depression likelihood ahead of the onset of illness, with different measures used: language, linguistic style, emotion, ego-network, demographics and user engagement.

We chose to tackle the problem of detecting early onset of depression from users' Reddit posts. To that end, we focus our efforts into processing the eRisk 2018 dataset (Losada et al., 2018), given its success at the Workshop for Early Risk Detection on the Internet[4] within The Conference and Labs of the Evaluation Forum (CLEF) and its fruitful submissions from participants.

The teams from this workshop had different detection systems, based on bag of words ensembles (Trotzek et al., 2018), machine learning models with hand-crafted features (Trotzek et al., 2018; Ramiandrisoa et al., 2018; Cacheda et al., 2018; Ramíırez-Cifuentes and Freire, 2018) or with different text embeddings (Trotzek et al., 2018; Ramiandrisoa et al., 2018; Ragheb et al., 2018), on sentence-level analysis to detect self references and extract different features (Ortega-Mendoza et al., 2018), on Latent Dirichlet Allocation (LDA) topic modelling (Maupomé and Meurs, 2018), models combining Term Frequency — Inverse Document Frequency with Convolutional Neural Networks (Wang et al., 2018) or other machine learning models. Most systems took the decision after the last chunk, only a few were able to emit a decision in the first chunks.

Several works addressing depression (Schwartz et al., 2014; Resnik et al., 2015) and PTSD (Coppersmith et al., 2015; Preoţiuc-Pietro et al., 2015) use a topic modelling approach showing that topics encountered texts have important discriminative power to make the distinction between persons suffering from mental illnesses and healthy controls.

3 Dataset

Early Risk Detection on the Internet (eRisk) workshops organized by CLEF explore the technologies that can be used for people's health and safety and the issues related to building tests collections (Losada et al., 2018). eRisk 2018 has two tasks, for early detection of depression and anorexia. We choose to focus on the task of detecting early onset of depression of social media users.

This task consists of sequentially processing chunks of Reddit posts from depressed users and controls. Submissions from each user are encoded in an xml file, one subject xml per chunk of data. Each xml contains the id of the subject and his posts and comments. Each submission has the posting time and the actual text. If a submission does

[4] https://early.irlab.org/

not have a title, it is considered a comment. The goal is to detect depression as early as possible and the dataset has to be processed in chronological order. The test collection of posts from depressed and non-depressed users is split into 10 chunks. As training data, the teams had access to data from eRisk 2017, both train and test. The test chunks were released one every week. Every week the teams had to decide whether to label the user as depressed or non-depressed or to wait for the test data of the following week.

The dataset contains 125 depressed users and 752 non-depressed users as training data and 79 depressed users and 741 non-depressed users as test data. The dataset has more posts and comments from people without depression than from users diagnosed with depression. From a total of 531,349 submissions, only 49,557 submissions are from users diagnosed with depression. The average time from the first to the last submission is between 2 and 3 years, so the posts were collected over a long period of time (Losada et al., 2018).

4 Method

Our methodology for early diagnosis of depression follows a classical Natural Language Processing pipeline. To clean the users' texts, we transform them into lowercase, we remove the punctuation and stopwords, the numbers and URLs are replaced with specific tokens and we perform stemming with Porter Stemmer (Porter, 1980). To reduce the dimension of the dictionary, we use collocations (Bouma, 2009) to extract meaningful bigrams and trigrams.

The number of posts and comments from non-depressed users is much higher than those from depressed users. To balance the two classes, we downsample the majority class to a ratio of 2:1.

We train our Latent Semantic Indexing model with 128 topics on every users' post. We use this model to extract topic modelling embeddings from users' texts and use them as input to our fully connected neural network architecture. The neural network has three hidden layers of 512, 256 and 256 neurons respectively, Leaky ReLU activation and we use Dropout for regularization. We use a random sample of 20% of the training data provided by the organisers of the competition for validation.

The network has two outputs, one for classifying if the user is depressed or not and one for confidence estimation. The motivation for using this architecture is to learn the confidence (DeVries and Taylor, 2018) of our predictions and use it to make a decision on whether to label a user or wait for the next chunk of data. The learned confidence, besides its use case in out-of-distribution detection, can be used as a measure for how much the model trusts its classification output to be correct. As such, we consider the classification output only if the confidence exceeds a certain threshold. As indicated by DeVries et al. (2018), the network loss is computed by interpolating the predicted probabilities p with the target y, using the computed confidence score c, as follows:

$$p_i' = c \cdot p_i + (1 - c)y_i \tag{1}$$

The final loss is then given by:

$$\mathcal{L} = -\sum_{i=1}^{M} log(p_i')y_i - \lambda log(c) \tag{2}$$

Where, in our case, $M = 2$, is the number of classes. The loss includes an additional term that forces the predicted confidence to be as high as possible. We performed an ablation study on the validation data on the confidence penalty λ.

A recent study by Hein et al. (2019) shows that neural networks with ReLU activation functions tend to be overconfident on incorrectly classified samples, thus we can not rely only on the output probabilities, and the predicted confidence offers a more reliable measure of uncertainty of the classification.

As the number of submissions seen by the model increases, we want to make a decision as early as possible and thus we use a decaying function that decreases progressively the fixed threshold for confidence. The decision function is defined as follows:

$$D_w(x) = \begin{cases} \text{decide for } x & \text{if } c > T * e^{-sw^2} \\ \text{wait for data} & \text{otherwise} \end{cases} \tag{3}$$

Where x is the embedding for the current user's posts, w is the week number (i.e. the current chunk), s is a scaling factor and T is the initial threshold. We choose $T = 85\%$ and progressively scale it down to 40%. The scaling factor is computed such that, at the final chunk, the threshold is less than the smallest confidence encountered on the training data.

At the test phase, the proposed model does not make an independent decision for each chunk of data in the test set. In the first chunk of data, if the model is not confident enough to make a final decision regarding the depressed or non-depressed status of a user, then, starting with the second chunk of data, we concatenate the current chunk with the previously available chunks for the current user. This way, the LSI model has more data for making better informed predictions.

5 Results

Our results on eRisk 2018 dataset are presented in Table 1. Even if F_1 is a standard evaluation measure used for imbalanced classification, it does not include the time component of the early detection task, thus Losada and Crestani (2016) propose an evaluation metric better suited for this task, the Early Risk Detection Error (ERDE).

ERDE is defined as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} \quad if \quad d = FP \\ c_{fn} \quad if \quad d = FN \\ lc_o(k) \cdot c_{tp} \quad if \quad d = TP \\ 0 \quad if \quad d = TN \end{cases} \tag{4}$$

The use of false positive (FP), false negative (FN), true positive (TP) and true negative (TN) for prediction *d* is to avoid the classifiers that always predict the label of the majority class. $lc_o(k) \in [0,1]$ encodes a cost for the delay in detecting TP. For the eRisk datasets, where the number of negative labels is greater than positive labels, the value of c_{fn} is 1 and c_{fp} is 0.1296, set according to the proportion of depressed users in eRisk 2017 dataset (Losada et al., 2018). c_{tp} is set to c_{fn} because the late detection of people at risk of depression can have serious consequences, a late detection is considered as equivalent to not detecting the depressed user at all. The late detection of TN cases does not affect the effectiveness of the system.

The goal of the system is to detect as early as possible people at risk of depression. For the detection of non-depressed users, the time of the detection is not relevant. The latency cost function, which grows with *k* (the number of submissions seen by the algorithm), is defined as:

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \tag{5}$$

o represents the number of posts after which the cost grows more quickly.

Method	$ERDE_5$	$ERDE_{50}$	F_1	Prec	Rec
Baseline LSI	**9.98%**	8.29%	0.25	0.22	0.29
LSI_c $\lambda = 0.01$	14.19%	11.41%	0.25	0.15	0.87
LSI_c $\lambda = 0.1$	11.12%	9.09%	0.28	0.20	0.48
LSI_c $\lambda = 0.2$	10.24%	**7.74%**	**0.30**	**0.25**	0.38
LSI_c $\lambda = 0.4$	11.15%	8.53%	0.25	0.17	0.47
LSI_c $\lambda = 0.6$	12.67%	10.17%	0.25	0.15	**0.71**
LSI_c $\lambda = 0.8$	10.53%	8.08%	0.30	0.21	0.56
Funez et al.(2018)	8.78%	7.39%	0.38	0.48	0.32
Trotzek et al.(2018)	9.50%	6.44%	0.64	0.64	0.65

Table 1: Classification results on the detection of early onset of depression task from eRisk 2018 dataset.

The detection task is difficult, as seen in the low values of F_1 and Precision. However, the task is to predict *early* onset of depression, and for that, the ERDE metrics are more appropriate, as they are a measure of prediction delay. $ERDE_5$ metric is very sensitive to delays, after the first 5 submissions from the user the penalties grow quickly. In contrast to $ERDE_5$, for $ERDE_{50}$ the penalties grow only after the first 50 submissions from the user. The difference between $ERDE_5$ and $ERDE_{50}$ is very important in practice because of the consequences of late detection of depression signs. As the task suggests, the detection should be made as early as possible.

To measure the impact of our learned out-of-distribution confidence from the neural network, we also trained a plain ReLU network with cross-entropy loss. For this model, we employed a hard threshold on the output probabilities for whether to wait for more data or classify the sample. As shown by Hein et al. (2019), ReLU networks can be overly confident on misclassified examples. This is shown in Table 1: the model has a low $ERDE_5$ score as the output probabilities mostly have extreme values, which means that for most users the model makes a decision from the first chunk of data.

We trained our model with different λ values in order to see the impact of the confidence component on the results. Larger values for λ make the model overly confident, as expected from Equation 2, the best performing model being the one with $\lambda = 0.2$. Smaller values of λ generate a wider confidence distribution on the training examples, facilitating the decision process, as extreme values either make the model overly-confident on every example, or not confident at all. This is consistent

with findings by DeVries et al. (2018).

In Table 1 we also present the best two submission from the eRisk 2018 Workshop, the one from Funez et al. (2018), having the best results for the $ERDE_5$ metric, and the one from Trotzek et al. (2018) having the top $ERDE_{50}$ score.

We can assume from these results that topics encountered in user writings have important discriminatory power. Depressed users mostly write about different subjects than non-depressed subjects, consistent with results from the work of Resnik et al. (2015). The writings from users diagnosed with depression are more focused on their feelings and their life events. Topics related to those themes contain words such as *someone kill, bad though, never able to get, forever alone, life save, stay sober, i am sad, still can't, improve life. new hope, oneself, tell anything, happy sad, hope one day*. Texts from non-depressed users are found in topics related to their hobbies containing specific words: *black mirror, first season, movie adaptation, hologram, nine inch nails, jimi hendrix, artist name, vlog, game, fallout, terra mistica, way to make money, paid time, really proud, amazon whishlist, food industry, white bread*.

6 Conclusion

In this paper, we use the eRisk 2018 dataset on Early Detection of Signs of Depression for depression classification from Reddit posts. Our method uses Latent Semantic Indexing for topic modelling and to generate the embeddings used as input for our neural network, but focuses on using a learned out-of-distribution confidence score alongside the classification output to decide whether to label the user or wait for more data. Besides its initial use case in out-of-distribution detection, we repurposed the confidence score as a measure for how much the model trusts its classification output to be correct. We showed that, in general, there is a significant difference in writing topics depending on the users' mental health, to the extent that it contains enough information for use in classification.

Acknowledgements

We would like to thank our reviewers for their useful comments and suggestions that helped us improve this paper and also to the organizers of the eRisk Workshop for their efforts in encouraging the research on mental illnesses detection from social media.

Liviu P. Dinu was supported by a grant of the Romanian Ministry of Education and Research, CCCDI—UEFISCDI, project number 411PED/2020, code PN-III-P2-2.1-PED-2019-2271, within PNCDI III.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Fidel Cacheda, Diego Fernández Iglesias, Francisco Javier Nóvoa, and Victor Carneiro. 2018. Analysis and experiments on early detection of depression. *CLEF (Working Notes)*, 2125.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Dario G Funez, Maria José Garciarena Ucelay, Maria Paula Villegas, Sergio Burdisso, Leticia C Cagnina, Manuel Montes-y Gómez, and Marcelo Errecalde. 2018. Unsl's participation at erisk 2018 lab. In *CLEF (Working Notes)*.

Sharath Chandra Guntuku, David Yaden, Margaret Kern, Lyle Ungar, and Johannes Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 12.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50.

Maria Khokhlova, Viviana Patti, and Paolo Rosso. 2016. Distinguishing between irony and sarcasm in social media texts: Linguistic observations. In *2016 International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–6. IEEE.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th International Conference of the CLEF Association, CLEF*.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87.

Diego Maupomé and Marie-Jean Meurs. 2018. Using topic extraction on social media content for the early detection of depression. *CLEF (Working Notes)*, 2125.

Taleen Nalabandian and Molly Ireland. 2019. Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or friends. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 62–73.

Rosa María Ortega-Mendoza, Adrián Pastor López-Monroy, Anilu Franco-Arcega, and Manuel Montes-y Gómez. 2018. Peimex at erisk2018: Emphasizing personal information for depression and anorexia detection. In *CLEF (Working Notes)*.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30.

Waleed Ragheb, Bilel Moulahi, Jérôme Azé, Sandra Bringay, and Maximilien Servajean. 2018. Temporal mood variation: at the clef erisk-2018 tasks for early risk detection on the internet. In *Proceedings of the 9th International Conference of the CLEF Association*.

Faneva Ramiandrisoa, Josiane Mothe, Farah Benamara, and Véronique Moriceau. 2018. Irit at e-risk 2018. In *Proceedings of the 9th International Conference of the CLEF Association*.

Diana Ramíirez-Cifuentes and Ana Freire. 2018. Upf's participation at the clef erisk 2018: Early risk prediction on the internet. In *Cappellato L, Ferro N, Nie JY, Soulier L, editors. Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum; 2018 Sep 10-14; Avignon, France.[Avignon]: CEUR Workshop Proceedings; 2018. p. 1-12.* CEUR Workshop Proceedings.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107.

Hannah Ritchie and Max Roser. 2018. Mental health.

H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125.

Ivan Sekulić, Matej Gjurković, and Jan Šnajder. 2018. Not just depressed: Bipolar disorder prediction on reddit. *arXiv preprint arXiv:1811.04655*.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF (Working Notes)*.

Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*.