UDE at eRisk 2019: Early Risk Prediction on the Internet

Razan Masood[1], Faneva Ramiandrisoa[2], and Ahmet Aker[1]

[1] University of Duisburg-Essen, Duisburg
{firstaname.lastname}@uni-due.de
[2] Université de Toulouse, France
faneva.ramiandrisoa@irit.fr

Abstract. In this paper, we describe our participation in CLEF eRisk workshop. eRisk 2019 is the third edition of this track [3], which was first introduced in 2017. In the current edition, the organizers are targeting Social Media users, namely Reddit, who may be under the risk of Anorexia, self-harm, and depression. We participated in both tasks of early risk detection of Anorexia and self-harm. Our predictions are based on Natural Language Processing using supervised machine learning with Support Vector Machines (SVM) and neural networks. SVM gave the best results among our five submitted models with *latency-weighted F1* of 0.58 and $ERDE_5$ of 0.08 and $ERDE_{50}$ of 0.04 for Anorexia detection task, while our more complicated neural network models did not show the desired performances.

Keywords: Early risk detection · Anorexia · Self-harm · NLP · LSTM · SVM.

1 Introduction

Social Media (SM) provided a more vibrant environment for communication and sharing experiences. However, different means of interactions on SM platforms induced enormous amounts of data of different variations. Feeding the data to machine learning and data mining algorithms revealed much of user personality that could not be known even by their close family members [4].

Hence, the users' mental health state is not an exception and might be revealed and understood through the use of SM data as well.

In mental health treatment, professionals use the textual content produced by people suffering from different kinds of mental illness to analyze and help them in the treatment and diagnosis. This procedure could be costly to do on

[3] http://early.irlab.org/

[4] https://www.nytimes.com/2015/01/20/science/facebook-knows-you-better-than-anyone-else.html

site, and it may not result in enough data. Therefore, mining social media could be a promising method to get more related content for analysis. Besides, it is not only a cheaper source but also, a reliable way to reach more people that can be at risk and in need of professional therapy [4, 5].

eRisk 2019 introduced three collaborative tasks to deal with the early risk detection of different mental illnesses, namely Anorexia, self-harm, and depression, using textual data from related Reddit subreddits. We submitted predictions for the first two tasks. We used five models for both of the tasks, including variations of SVM and LSTM neural networks. In the following, we introduce our models and results.

2 Data

The data was extracted from Reddit[5]. This website includes many communities to discuss certain topics and interests in what are called *subreddits*.

The first task is related to the early detection of signs of Anorexia. This task has two stages training and testing. Training and Test datasets of the same task in eRisk 2018 were joined to be used as the training set for eRisk 2019. The data consists of users' messages, which are separate posts with a title and a textual content.

In earlier anorexia and depression detection challenges, users' posts were divided into ten chunks each with 10% of each user posts. In 2019 anorexia challenge, the test data is provided in an item-by-item manner. The predicting model must submit a decision for the user, i.e., positive or not, upon receiving each message from their stream. A score indicating the level of Anorexia must also be submitted after receiving each writing.

The second task is related to the early detection of Signs of Self-harm. The source of the data is the same as the first task with the same test stage format. There is no training data for this task, so it is presented as an unsupervised learning task to motivate search based methods.

More information about the data is provided by [8].

3 Methodology

We first pre-processed the textual data by lemmatizing, lower casing, and then removing stop words, new-lines, and other unrelated terms or symbols related to the Reddit platform. After the cleaning process, we used different machine learning approaches to address the tasks. In our strategy, we aimed to use methods that are as simple as possible, but that led in earlier cases to competitive results [10] and then also try out more complex models that showed successes in related areas [12].

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

3.1 Linear Support Vector Classification

For the first model, we employed a traditional machine learning (ML) model, which is an SVM with a linear kernel. For the training phase, we concatenated all the writings for each labeled user, including text and title parts. Then, we applied a term-frequency transformer and then performed feature selection using *chi-squared* test to select the most significant 500 terms. On the resulting feature vectors from the described pipeline, we performed a 10-fold cross validation using *LinearSVC* provided by Python *sklearn* library. We used *grid search* to determine the SVM model parameters. Besides, we defined a class weight to overcome the unbalanced data problem. The parameters that were used are: $\{penalty = "l2", dual = True, tol = 1e-3, C = 10, class\_weight = 1:4\}$.

3.2 Filtered Linear Support Vector Classification

This method is a two-steps classification. The idea is based on the fact that by observing each user's posts, we found that the stream contains many posts that are not related to the mental health issue and discuss other topics like games, movies or politics. Hence, the first step is to filter out the writings that have a higher probability of being not related to the topics discussed frequently by mental health subreddit users and keep only those that are more likely to be about mental illness related issues. These most related writings pass to the next classification step to predict if the user is under risk or not. If no writings pass this filter, the user's label remains negative. In this model, we tried to narrow down the writings to get the ones that have a higher probability of having informative features on the users' mental health and examine if this will enhance the SVM performance by adding a pre-filtering phase.

To perform the first step, we manually annotated writings in the training set that are not related to mental health. Our guidelines were to filter out any post that does not discuss any of the following issues: depression, food or cooking, physical activity, medications, life experience/story, self-harming/suicidal and eating disorder experiences. These topics are selected based on literature that mentioned topics written by people with mental health issues and by manually observing positive users postings in the data [11, 13]. Two people have performed the data annotation. We first chose randomly 200 posts and annotated them. The inter-rater agreement computed using Kappa on these 200 posts is $\kappa = 0.72$.

We then extended the dataset and annotated more posts. In total, we have 660 posts. From this 660 posts, 230 (34.85%) are related to mental health, and 430 (65.15%) are not. We trained a linear SVM with similar features to the one trained for the first model described in Section 3.1, but we selected the most significant 200 terms. We trained the SVM using 10-fold cross-validation on the 660 writings. The best parameter combination, found using grid search, produced 0.68 as the mean cross-validated *F1* score. These parameters are: $\{tol = 1e-3, C = 1, class\_weight = "balanced", gamma = "scale"\}$.

We consider this described classifier as a filter which does not allow the writing to pass to the next step if the probability of it not being related to

mental health is higher than 95%. We chose this high probability to reduce the loss of true positives.

The second step classifier is responsible for the writings that pass through the previous filtering step. It is the same classifier described in Section 3.1. To train this classifier, we filtered the training data using the step-1 model and used the resulted data for training.

3.3 LSTM Model

This model is based on a vanilla LSTM (Long Short Term Memory) neural network, which is a type of recurrent neural network.

Since the eRisk data is a stream of user's posts ordered by time, LSTM could be applicable for classifying users. For training this model, we did not use the full posts stream, but we used only 45 writings of the users' streams by which we took 15 writings from the beginning, another 15 from the middle and another 15 from the end of the stream. We chose the number of writings based on manual observation of the data in order to summarize the users' writing stream. The goal is to have a minimum and a representing number of writings to asses the users' risk. The features extracted from these writings are a concatenation of *doc2vec* features and term frequency features. We trained the model on the concatenation of all writings (including title and text) for each user.

The *doc2vec* model produces a 200 long vector which is a concatenation of the results of two *doc2vec* algorithms, namely *Distributed Bag of Words* and *Distributed Memory* with an output of 100 long vector each and trained in the same way described by [14] using Python implementation provided by *gensim* library. We concatenated the 200 *doc2vec* feature vector with a vector of term frequencies of the most significant 70 terms selected as described in Section 3.1.

We used *Keras* implementation of LSTM with 500 units for output, a dropout of 0.2, a dense layer of size 2 and *Softmax* function for the final output. We used *binary_cross-entropy* for the loss function and *Adam* as an optimizer. We defined *F1* metric for model evaluation.

In the test stage, the input to the trained model is given one writing at a time. Whenever new writings arrive, we concatenate it with the previous ones and use the concatenation as input to the model. When the stream of writings became longer, we started to arrange the input to concatenate only 45 writings the way we described above.

3.4 Global Attention Model

Attention is a mechanism used in deep learning models that have been quite popular lately, and first appeared in neural machine translation. *Attention* was mainly introduced to address the inefficiency of sequence-to-sequence encoding in memorizing longer sentences [1]. This mechanism allows the model to learn for each word in the target sequence which words to attend to (pay attention to) in the input sequence by learning alignment weights between that pairs of

output and input words. These alignments are in turns used to calculate the final context vector for each word/ time-step. In our case, time steps are the writings in the user's stream. Accordingly, the model is supposed to learn the importance of each of the writings to predict if the owner is at risk of Anorexia.

The model's input is similar to what we described for the previous simple LSTM model in Section 3.3 but using only the *doc2vec* feature vectors, i.e. a *doc2vec* for each writing. The 45 vectors formed an input layer for another LSTM layer with 16 units and then followed by an attention layer. We use what is called global/ soft attention, as described by [9], which is a simplification of the attention mechanism in [1]. We use a dropout of 0.1 and then normalizing the attention output with the *Softmax* function and predict a positive result when its probability is higher than 0.5. In the test phase, we used the same strategy described in Section 3.3.

3.5 Inner Attention Model

By manually comparing writings streams of positive labeled users and negative labeled ones, we noticed that the positive users' writings contain different kinds of topics and information that is more frequent in their feed. These topics, for example, can be related to diet, eating habits like fasting and food, and medications. Hence, the idea is to train a model that learns the importance of each writing depending on its topic.

This idea is similar to the model proposed in [12] for arguments classification. In their paper, Stab et al. use an average vector of embedding of each topic's terms to build the attention layer. These topics embeddings allowed to detect whether a sentence constitutes an argument or not by engaging the topic of the sentence in the detection model. For our case, we used *doc2vec* of paragraphs collected from web sources related to Eating Disorders in addition to posts from ED related *r/EatingDisorders/* and *r/AnorexiaNervosa/* subreddits, instead of topic terms that were used in the original paper. We used collected paragraphs from eight different web sources that contain articles about Anorexia and other eating disorders[6,7,8]. We used 15 paragraphs from the ED related websites chosen manually to be as diverse as possible to consider different aspects of the disorder. Besides, we added 15 writings from the mentioned subreddits. We selected the writings to cover different topics that users mention in these subreddits, such as body/weight information, diets, and recovering experiences.

The implemented model includes an inner-attention layer that receives both the input as *doc2vev* features from 45 writings as described before and the doc2vec feature of 30 paragraphs of related topics and an LSTM with 64 units with a dropout of 0.1 and a dense layer of 2 units for the output using *Softmax* function.

[6] https://www.mayoclinic.org/diseases-conditions/anorexia-nervosa/symptoms-causes/syc-20353591

[7] https://www.psycom.net/eating-disorders/anorexia/

[8] https://www.eatingdisorderhope.com/information/eating-disorder

4 Results

4.1 Task 1: Early Detection of Signs of Anorexia

We developed the five models described in the earlier sections for Task 1, namely Anorexia detection. eRisk organizers evaluate the prediction based on *F1*, precision, recall, and $ERDE$ measure, which was first proposed in [6]. However, $ERDE$ measure has some deficiencies. Hence, eRisk 2019 introduced *latency-weighted F1* score measure [8]. The linear SVM model produces the best *latency-weighted F1* score among our other four submitted models. See Table 1 for results. Whereas, other NN models performed poorly on this task. The SVM model ranked 11th between 13 teams' 54 submitted models according to the *latency-weighted F1* measure.

Table 1. Results of the five models in task 1: Anorexia detection

Model	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	*latency-weighted F1*
SVM	.51	**.74**	.61	.08	.04	**.58**
Fitered-SVM	.44	.73	.55	.07	.04	.53
LSTM	.13	.68	.22	.13	.08	.19
Global-attention	0	0	0	-	-	-
Inner-attention	0	0	0	-	-	-

The organizers have added new ranking-based evaluation measures that are based on the submitted scores that accompanied each received writing. The goal is to rank the users according to their estimated risks [8]. The standard IR metrics *P@10* and *NDCG* were reported after seeing 1, 100, 500 and 1000 writings. According to this evaluation metric, the filtered-SVM model performed slightly better than the single step SVM model. Again, the other NN models performed poorly according to this measure. Results are shown in Table 2. Nonetheless, these measure values were high compared to other teams, but this could be a result of a relatively high recall of our models.

Table 2. Ranking-based evaluation of the five models in task 1

Model	1 writing			100 writings			500 writings			1000 writings		
	P@10	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*
SVM	.2	.12	.11	.9	.92	.81	.9	.93	.85	.9	.94	.86
Fitered-SVM	.6	.75	.54	.9	.94	.81	1	1	.87	1	1	.88
LSTM	.7	.76	.49	.9	.94	.60	.9	.94	.64	.8	.88	.64
Global-attention	.0	.0	.11	.0	.0	.08	.0	.0	.06	.0	.0	.07
Inner-attention	-	-	-	-	-	-	-	-	-	-	-	-

4.2 Task 2: Early Detection of Signs of Self-harm

Many studies and investigations showed that different mental illnesses accompany and relate to each other. For example, people who suffer from Anorexia

could be depressed or self-harming and vice versa [3, 2]. Since task 2 had no training data, we wanted to investigate if training the model on writing provided for different tasks, namely depression detection, and anorexia detection could help in detecting users at risk of self-harm. Therefore, we trained the same five classifiers described before for Anorexia detection on previously provided eRisk data for depression detection in addition to the data provided for Anorexia detection. Precisely, we used the datasets provided by eRisk 2018 [7] on depression and Anorexia and used the positively labeled users in each as positive self-harm cases. The performance was poor using the mentioned datasets for training according to the measures proposed by the organizers which were mentioned in the previous section. See Table 3. The ranking evaluation indicated higher performance for this task as well. See Table 4.

Table 3. Results of the five model in task 2: Self-harm detection

Model	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	*latency-weighted F1*
SVM	.50	.07	.13	.12	.11	.12
Fitered-SVM	.45	.22	.30	.11	.10	.29
LSTM	.18	.68	.29	.14	.10	.28
Global-attention	0	0	0	-	-	-
Inner-attention	.06	.34	.10	.20	.20	.07

Table 4. Ranking-based evaluation of the five models in task 2

Model	1 writing			100 writings			500 writings			1000 writings		
	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
SVM	0	0	0.09	.7	.77	.69	.7	.67	.69	.7	.67	.70
Fitered-SVM	.7	.56	.52	.7	.66	.69	.8	.75	.74	.8	.75	.74
LSTM	.5	.63	.53	.5	.56	.64	.6	.66	.68	.6	.65	.67
Global-attention	.2	.25	.30	.1	.07	.20	.1	.07	.15	.1	.08	.17
Inner-attention	.2	.19	.21	0	0	.11	.2	.16	.14	.1	.07	.15

5 Conclusions and Future Work

In this paper, we introduced the working notes of our participation in eRisk 2019 for the early risk detection of Anorexia and self-harm. We proposed solutions based on traditional ML with linear SVM and NNs, including LSTM and attention mechanisms. What makes this task tricky is the trade-off between deciding earlier with a low number of posts that a user is at risk of having the negative consequences of mental illnesses symptoms on the one hand, and being wrong to classify such user to be at such risk. Accordingly, by looking at the data and trying to decide whether a user is at risk by reading their posts one by one, it seemed hard to decide with much confident especially when some users who are already labeled as positive cases do not have many posts. However, introducing

a score to guide the level of risk and ranking users according to their estimate of risk could be a more reliable indicator to use, as stated by [8].

SVM did not perform as it is expected, this could be due to the difference between the training and testing settings, where the training is done using the stream of the whole writing for each user, whereas the judging is associated with each writing at a time. On the other hand, what is most tricky by this kind of problem is the coding for a user's writing stream. The NN solutions we used are initially used for a stream of words that constitute a sentence that needs to be classified, but in our case, we have several paragraphs/sentences that belong to one user which should be classified. It is most likely that our input encoding did not fit with the test and evaluation strategy of the tasks. On the other hand, as mentioned before, the performance was better in terms of ranking evaluation when using more writings.

In our future work, we aim to investigate methods that suit the item-by-item test phase. Moreover, we need to investigate better encoding for the user input without losing much knowledge by encoding each post as one unit of information.

6 acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group User-Centred Social Media.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Benton, A., Mitchell, M., Hovy, D.: Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538 (2017)
3. Coopers, P.: The costs of eating disorders: Social, health and economic impacts. B-eat, Norwich (2015)
4. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in twitter. In: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality. pp. 51–60 (2014)
5. Diederich, J., Al-Ajmi, A., Yellowlees, P.: Ex-ray: Data mining and mental health. Applied Soft Computing **7**(3), 923–928 (2007)
6. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Conference Labs of the Evaluation Forum. pp. 28–39. Springer (2016). https://doi.org/10.1007/978-3-319-44564-9_3, http://dx.doi.org/10.1007/978-3-319-44564-9_3
7. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2018: Early risk prediction on the internet (extended lab overview)
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019. Springer International Publishing, Lugano, Switzerland (2019)

9. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
10. Ramiandrisoa, F., Mothe, J., Benamara, F., Moriceau, V.: Irit at e-risk 2018. In: E-Risk workshop. pp. 367–377 (2018)
11. Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: 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. pp. 99–107 (2015)
12. Stab, C., Miller, T., Gurevych, I.: Cross-topic argument mining from heterogeneous sources using attention-based neural networks. arXiv preprint arXiv:1802.05758 (2018)
13. Toulis, A., Golab, L.: Social media mining to understand public mental health. In: VLDB Workshop on Data Management and Analytics for Medicine and Healthcare. pp. 55–70. Springer (2017)
14. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression. In: CLEF (Working Notes) (2017)