TUA1 at eRisk 2018

Ning Liu*, Zheng Zhou*, Xin Kang, and Fuji Ren

Tokushima University, Tokushima 770-8506, JP
liuning08@gmail.com
c501837040@tokushima-u.ac.jp
{kang-xin, ren}@is.tokushima-u.ac.jp

Abstract. Mental health is important at every stage of life in the modern world. The eRisk gives two tasks on detection of depression and anorexia respectively. In this paper, we do not take any temporal information or other corpus to support the tasks. We employ a TF·IDF with SVM as classifier, a CNN+LSTM based deep neural network and a simple keywords based method to very whether those methods can learn the mental information from sparse space with unbalanced small data sets. The task results show the simple keywords model gets the best results in task 1 of f1 0.47 and the best recall of 0.71. In task 2, the simple keyword model gets the best recall score of 0.76, CNN+LSTM model gets the best f1 score of 0.36.

Keywords: Early risk detection · feature representation · CNN · LSTM.

1 Introduction

Mental health is important at every stage of life in the modern world. According to WHO, globally, more than 300 million people suffer from depression, the leading cause of disability. More than 260 million are living with anxiety disorders. Many of these people live with both[13]. The mental disorder not only hurts the people themselves, but also causes harm to their friends, families or even end someone's life. With the fast-paced life and pressed working environment, people are easy to get unhealthy mental status. For human who are suffering from these hurts, without professional intervention, they cannot get through by themselves. To be worst, suffering the mental health for someones always means no communication with others. Without diagnosis, the doctors cannot give any detections.

Since we are in the Internet era, many people submit their comments, texts, photos or videos to the social networks. For the people with mental health problems, it is easier to analyze their mental situation through the memories written in the social networks. CLEF 2018 gives two sub tasks in eRISK(early risk prediction on the Internet) task[8]. One is early detection of signs of depression, the other one is early detection of signs of anorexia. The source data is crawled from a set of social media users, and are formatted using the collection described in

* These authors contributed equally to this work

[6]. The data in every subtasks are split into ten chunks, each chunk contains 10% of the messages. The earlier chunk detected, the higher ERDE score got.

As we took part in the task from April, we don't have enough time to train complex models. To the time sequences prediction problems, temporal feature is one of the important aspects to be considered, external corpus or more complex semantic models are also efficient. In this paper, we deal with this task in simple ways to explore whether simple models or some key features will work for this time correlation task without special temporal features considered. To achieve these goals, we construct a traditional CNN+LSTM based deep neural network, employ TF·IDF represented features on SVM model and specific keywords selected method respectively. We only submit the tenth chunk, and the final chunk scores show some exciting results.

The remainder of this paper is organized as follows: Section 2 gives related works. The three verified method will proposed in Section 3. Section 4 presents the results. Section 5 makes the conclusion and future work.

2 Related Work

Emotion recognition related tasks have been held for years in different conferences or workshops. For example, sentiment analysis in Twitter[11] of SemEval-2017, emotion cause analysis in NTCIR-13[4] and early risk prediction on the Internet in CLEF 2017[7]. These tasks are held in variant languages:English, French, Chinese, Arabic and so on. Kawachi et al.[5] describe the social categories and the related social relationships of people with mental health outputs, their research shows the social ties play a beneficial role in the maintenance of psychological well-being. College students are surfing more and more depression in social media or Internet[9]. And they want to cue themselves firstly by communication with other in the social networks[12]. Thus psychology researches exam the behavioral characteristics of depression, anorexia or other mental disorder activities[14]. The characterization study shows young individuals have two prominent anorexia related communities on Tumblr– pro-anorexia and pro-recovery and provides an empirical analyses on several thousands Tumblr posts[3].

3 Models

The eRisk 2018 contains two sub-tasks, one is Task 1: Early Detection of Signs of Depression, and the other one is Task 2: Early Detection of Signs of Anorexia. The tasks are both running on the contents crawled from social networks with temporal tags. Our team takes part in the two sub-tasks, we propose three methods to detect the early signs of mental disorder.

Keywords model Marked as TUA1B in Task 2 and TUA1C in Task 1 is a simple model in which using some emotional words as targets to measure the authors' situation. Research in [3] shows pro-anorexia community uses microblogging platform to share image-rich graphic and ”triggering” content around internalization of thin body ideals. According to this empirical description, we select the measurement keyword of anorexia as ”body”, and the strategy for measurement is if the contents of specific chunks contain the keyword, we will give the conclusion for this sample. The same way for depression detection, we give ”depression” as the keyword.

TF·IDF with SVM Marked as TUA1D in Task 1 and TUA1A in Task 2. This is a traditional method to assess the situation of mental health. We train the SVM model using a linear kernel API of sklearn[10] upon the full chunks contents. All of the texts in the chunks are processed into feature vectors using TF·IDF package in sklearn. The TF·IDF scores of every post is normalized under l2 function. The same strategy used above: samples detected will not consider in next chunks.

CNN+LSTM model Marked as TUA1A and TUA1B in Task 1, TUA1C in Task 2. We construct a CNN+LSTM based deep neural network as the detection models which can be seen in Table 1 using keras[2] with TensorFlow[1] as backend.

Table 1. The structure of CNN+LSTM model

Layer	Output Shape	param #
Embedding	(None, 2000, 128)	20043904
Dropout	(None, 2000, 128)	0
Conv1D	(None, 1996, 64)	41024
MaxPooling1D	(None, 499, 64)	0
LSTM	(None, 70)	37800
Dense	(None, 1)	71
Activation	(None, 1)	0
Total params: 20,122,799		
Trainable params: 20,122,799		
Non-trainable params: 0		

In this model, the contents of every chunk are preprocessed into one-hot features with an index value of the corresponding word in vocabulary instead of the word itself. Before feeding into the network, adding a padding process to format the length, the ”maxlen” is selected as 2000. Other hyperparameters chosen for this model are as follows: ”input_length” of Embedding is set to the length of vocabulary, Task 1 is 429700 and Task 2 is 156593, ”embedding_size” is 128; The dropout factor is 0.25; For convolution layer, the ”filters” is 64, setting ”kernel_size” to 5, ”padding” with ”valid”, ”strides” with 1 and ”activation” is ”relu”; Giving MaxPooling the parameter of 4 as ”pool_size”; For LSTM layer, the output dimension is 70; The following Dense layer is a one cell classification layer with ”Activation” being ”sigmoid”. For the compiling, choosing ”binary_crossentropy” as ”loss” and ”adam” for ”optimizer”. The ”metrics” is the default setting as ”accuracy”. Epoches is added as 20 to maximize the accuracy and minimize the loss of training data. In our experiments, the final loss in Task 1 is 0.000108512764422, in Task 2 is 0.00640664884888 respectively and accuracy are both 1.0.

4 Results

4.1 Data prepossessing

The crawled contents for two sub-tasks are formatted in XML files. Data sets are divided into ten chunks under posting timeline. For Task 1, the data sets contain training and testing sets of 2017 and this year's data for testing. In Task 2, the files only contain training data and testing data, cause this sub-task is a new task in eRisk this year. Contents for every IDs are organized with five tags: one root tag named $< WRITING >$ means one formatted posts, four child tags named $< TITLE >$, $< DATA >$, $< INFO >$ and $< TEXT >$ orderly.

$< TITLE >$ tag gives the title of this post, $< DATA >$ tag marks the post data, $< INFO >$ tag reminds which platforms the posts are crawled and the final $< TEXT >$ tag contains the contents people write-in. As $< TEXT >$ tag and $< TITLE >$ tag may both have no contents in one $< WRITING >$, to gain more textual information, during processing, we combine both $TITLE$ and $< TEXT >$ contents together, and extend the texts to one sentence. Although most of the IDs have a lot of $< WRITING >$ tags with efficient contents in one chunk, there are still IDs with none content in both $< TEXT >$ tag and $< TITLE >$ tag of one chunk and cannot extract useful information except temporal information (not used in this paper). At this moment, the chunk with none content can only contain the texts from former chunks. All of the contents of IDs are extracted into ten chunks, in which chunk1 only contains the current texts of this chunk, the next chunk can contain the contents of former one and contents contained in current chunk, following this way, we get split contents for chunk1, chunk2,$\cdots$,chunk10. Specially, for Task1, the training and testing data sets of 2017 are combined into one data set as new training corpus.

For the risk detection using keyword model, the $< TEXT >$ tag and $< TITLE >$ tag combined sentences of IDs will be very chunk by chunk without summing them together.

4.2 Evaluation results

Employing the models mentioned above with the extracted data, we submit the chunk 10th results of Task1 and Task 2 respectively. Table 2 and Table 3 show the results of Task 1 and Task 2 respectively.

Table 2. Results of four models in Task 1

Models	$ERDE_5$	$ERDE_{50}$	F1	P	R
Kerwords model(TUA1C)	10.86%	9.51%	**0.47**	**0.35**	**0.71**
TF·IDF with SVM(TUA1D)	0	0	0.00	0.00	0.00
CNN+LSTM(TUA1A)	10.19%	9.70%	0.29	0.31	0.27
CNN+LSTM(TUA1B)	10.40%	9.54%	0.27	0.25	0.28

Table 3. Results of three models in Task 2

Models	$ERDE_5$	$ERDE_{50}$	F1	P	R
Kerwords model(TUA1B)	19.90%	19.27%	0.25	0.15	**0.76**
TF·IDF with SVM(TUA1A)	0	0	0.00	0.00	0.00
CNN+LSTM(TUA1C)	13.53%	12.57%	**0.36**	**0.42**	0.32

Our final chunk results of three models can be checked in Table 2 and Table 3. In Task1, the keywords model gets the best F1, precision and recall scores of 0.47, 0.35 and 0.71, two CNN+LSTM models get almost the same results. The two CNN+LSTM models are just two times running feeds with the model. The recall keywords model gets is one of the top recalls in the 11 teams, in which the best is 0.95. In Task 2, we employ three models, the keywords model only gets the best recall of 0.76, CNN+LSTM model get the best F1 and precision scores of 0.36 and 0.42.

TF·IDF with SVM gets both 0 value in Task 1 and Task 2, this makes us confused. This model with l2 normalization of TF·IDF features predict all the IDs as label ”2”, no risk ID detected in the tenth chunk. Those results make the model unbelievable, thus gets the 0 evaluation results finally.

Table 4. Ten chunks results of keywords model in Task 1

Evaluations	chunk1	chunk2	chunk3	chunk4	chunk5	chunk6	chunk7	chunk8	chunk9	chunk10
$ERDE_5$	9.87%	9.95 %	9.96 %	10.03%	10.17%	10.23%	10.29 %	10.35%	10.52%	10.55%
$ERDE_{50}$	9.63%	9.64 %	9.19 %	8.87%	8.90%	8.70%	8.58 %	8.40%	8.47%	8.02%
F1	0.31	0.37	0.41	0.43	0.42	0.44	0.46	0.46	0.45	0.47
P	0.47	0.44	0.44	0.42	0.38	0.38	0.38	0.37	0.34	0.35
R	0.23	0.32	0.39	0.44	0.48	0.53	0.58	0.61	0.65	0.71

As we only submit the tenth chunk results, for a fully analysis of the models, we calculate the ten chunks results of $ERDE_5$, $ERDE_{50}$, F1, P and R for keywords model in Task 1 and task2, which shows in Table 4 and Table 5 separately.

In Task 1, checking Table 2 and Table 4, we can find keywords model get the best F1 scores from the first chunk, while in Task 2, comparison the Table 3 and Table 5, The keywords model perform worse than in Task 1, though its recall

Table 5. Ten chunks results of keywords model in Task 2

Evaluations	chunk1	chunk2	chunk3	chunk4	chunk5	chunk6	chunk7	chunk8	chunk9	chunk10
$ERDE_5$	15.78%	16.90%	17.50%	17.82%	18.10%	18.62%	19.06 %	19.26%	19.58%	19.90%
$ERDE_{50}$	15.78%	16.88%	15.71%	15.32%	15.60%	16.03%	16.25 %	16.29%	16.22%	15.98%
F1	0.27	0.28	0.27	0.27	0.27	0.26	0.26	0.25	0.25	0.25
P	0.20	0.18	0.18	0.17	0.17	0.16	0.16	0.15	0.15	0.15
R	0.44	0.56	0.61	0.63	0.66	0.68	0.71	0.71	0.73	0.76

still perform good than CNN+LSTM method. Considering CNN+LSTM model only, with less negative samples CNN+LSTM model gets higher results in Task 2 compared with the two years data of Task 1. This maybe the small scale of the data sets cannot fully train the deep neural network, with more negative data, the model may learn more unbalanced information.

For keywords model, the keywords used are "depression" and "body" respectively for Task 1 and Task 2. They both get the best recall scores, While "depression" gets better results than "body" in F1 and Precision. These may indicate that in depression risk signs, people are often using the obvious words like "depression" to express themselves, while in anorexia situation, the disorder of eating may not almost focus on "body", there are maybe other hidden words existing. In other words, depressed people tend to express in the social networks directly and anorexia people are more difficult to detect. The team of FHDO gets the best detection among all the teams of F1 score of 0.85, we are looking forward to see their perfect models to be published.

5 Conclusion and Future work

In this paper, we propose three models to detect the risk of depression and anorexia respectively. Our results show the CNN+LSTM and keywords model can efficiently cover these task, while SVM model got unbalanced training. Our experiments show with more negative data sets, the CNN+LSTM model become sensitive about the data and need more tuning to train a suitable model for depression task. Our keywords model get the most efficient recall scores among the three models, even gets a good position among all teams, that proofs the key information is important to deal with these tasks.

In the future, we will use the test data to tune the CNN+LSTM model for a better detection ability combined the keywords information. Further more, a basic model by SVM without l2 normalization will also be checked.

Acknowledgments

This research has been partially supported by JSPS KAKENHI Grant Number 15H01712.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
3. De Choudhury, M.: Anorexia on tumblr: A characterization study. In: Proceedings of the 5th International Conference on Digital Health 2015. pp. 43–50. ACM (2015)
4. Gao, Q., Hu, J., Xu, R., et al.: Overview of ntcir-13 eca task. In: Proceedings of the 13th NTCIR Conference. Tokyo, Japan (2017)
5. Kawachi, I., Berkman, L.F.: Social ties and mental health. Journal of Urban health **78**(3), 458–467 (2001)
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), http://dx.doi.org/10.1007/978-3-319-44564-9_3
7. Losada, D.E., Crestani, F., Parapar, J.: Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations
8. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk – Early Risk Prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). Avignon, France (2018)
9. Moreno, M.A., Jelenchick, L.A., Egan, K.G., Cox, E., Young, H., Gannon, K.E., Becker, T.: Feeling bad on facebook: Depression disclosures by college students on a social networking site. Depression and anxiety **28**(6), 447–455 (2011)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
11. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation. SemEval '17, Association for Computational Linguistics, Vancouver, Canada (August 2017)
12. Vedula, N., Parthasarathy, S.: Emotional and linguistic cues of depression from social media. In: Proceedings of the 2017 International Conference on Digital Health. pp. 127–136. ACM (2017)
13. WHO: World mental health day 2017. http://www.who.int/mental_health/world-mental-health-day/2017/en/ (2017)
14. Wolf, M., Sedway, J., Bulik, C.M., Kordy, H.: Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. International journal of eating disorders **40**(8), 711–717 (2007)