# A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text

Yu-Tseng Wang[1], Hen-Hsen Huang[1], and Hsin-Hsi Chen[12]

[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
{ytswang,hhhuang}@nlg.csie.ntu.edu.tw, hhchen@ntu.edu.tw

**Abstract.** In recent years, people actively write text messages on social media platforms like Twitter and Reddit. The text shared on social media drives various applications including influenza detection, suicide detection, and mental illness detection. This work presents our approach to early risk detection of depression and anorexia on social media in CLEF eRisk 2018. For the two mental illnesses, our models combine TF-IDF information and convolutional neural networks (CNNs) to identify the articles written by potential patients. The official evaluation shows our models achieve $ERDE_5$ of 10.81%, $ERDE_{50}$ of 9.22%, and F-score of 0.37 in depression detection and $ERDE_5$ of 13.65%, $ERDE_{50}$ of 11.14%, and F-score of 0.67 in anorexia detection.

**Keywords:** Early Risk Detection, Depression, Anorexia, Convolutional Neural Network.

## 1 Introduction

In this work, explore people sharing their opinions, experiences, and feelings, on social media platforms from Twitter and Reddit. Textual information extraction can be used for various intelligent applications in the real world such as healthcare, communication, entertainment, journalism, and advertising. According to data from 2010 to 2018 reported by statista.com[1], the number of Facebook users increased from 431 million to 2,234 million, and the number of Twitter users grew from 30 million to 330 million. As of April 2018, Reddit had about 33 millions of users. In social media, life experiences and conversation history from a large number of users are recorded. In recent years, there is a variety of research focused on social media, including hate speech detection [1], information extraction [2], analysis on gender differences [3], nastiness detection [4], named entity recognition [5].

In most cases, the detection task can be considered as a classification problem. Various learning models and linguistic features are explored to deal with different goals. For example, the detection of terrorist attack needs to take latency into account because it is extremely important to prevent an attack from happening. Similar situations also

[1] https://www.statista.com

hold in the detection of illnesses. In CLEF eRisk 2018[2], two tasks on early risk detection of mental illnesses are conducted. The goal is to find out potential patients of depression and anorexia as early as possible. In other words, we aim not only to accurately predict if a social media user is a patient of depression/anorexia, but also to minimize the revealed user information. In contrast to usual detection tasks, early risk detection is more challenging. In this work, we conduct an analysis on the datasets and propose a neural network-based approach to the two detection tasks. The rest of this paper is organized as follows. Section 2 briefly describes the CLEF eRisk 2018 task and the dataset. We present our model in Section 3. In Section 4, experimental results are discussed. Section 5 concludes this work.

## 2 CLEF eRisk 2018 Task

### 2.1 Task Description

Early risk prediction on the Internet (eRisk), which started since 2017, is a task held in the Conference and Labs of the Evaluation Forum (CLEF) based on the consideration that automatic detection models could be applied to identify the risk as early as possible to help people avoid becoming victims of mental illnesses. In eRisk 2017 [6], a pilot task on the detection of depression is conducted, and the metrics including precision (P), recall (R), F1-score, and Early Risk Detection Error (ERDE) [7] are used for evaluation.

In this year, eRisk 2018 extends eRisk 2017 by introducing another mental illness, anorexia, to detect. In addition, the dataset of depression detection is also extended. Both tasks are organized in training stage and test stage. The training data is the writing history of users who are labeled as either risk or safe. The test data is composed of ten chunks released sequentially. For each chunk of a user's data, the model has to make a decision among three choices: (1) The model does not want to emit a decision on this user in this time. (2) The model emits a risk on this user. (3) The model emits a non-risk on this user. In Chunk 10, the last chunk, the undecided users should be determined as either risk or non-risk.

### 2.2 Datasets

In eRisk 2018[8], the datasets on depression and anorexia are released. Table 1 shows the statistics of the training sets. The posts and comments on Reddit, submitted by normal and risk users, are collected. In both datasets, we observe that the average submission per user in the normal group is higher than that in the risk group. On the other hand, the average length per submission in the normal group is lower than that in the risk group. Compared with Table 2, where the statistics of the test sets are shown, similar phenomena are also observed.

[2] http://early.irlab.org/index.html

**Table 1.** Statistics of the training sets.

| | Depression | | Anorexia | |
|---|---|---|---|---|
| | Risk | Non-Risk | Risk | Non-Risk |
| Number of subjects | 135 | 752 | 20 | 132 |
| Number of submissions | 49,557 | 481,837 | 7,452 | 77,514 |
| Submissions per subject | 367.1 | 640.7 | 372.6 | 587.2 |
| Words per submission | 27.4 | 21.8 | 41.2 | 20.9 |

**Table 2.** Statistics of the test sets.

| | Depression | | Anorexia | |
|---|---|---|---|---|
| | Risk | Non-Risk | Risk | Non-Risk |
| Number of subjects | 79 | 741 | 41 | 279 |
| Number of submissions | 40,665 | 504,523 | 17,422 | 151,364 |
| Submissions per subject | 514.7 | 680.9 | 424.9 | 542.5 |
| Words per submission | 27.6 | 23.7 | 35.7 | 20.9 |

The words with the highest TF-IDF score in the risk and the normal groups in both datasets are listed in Table 3. The top words of the anorexia patients, marked as bold, denote cues to the illness.

**Table 3.** Top 20 words with highest TF-IDF score from test data.

| Rank-ing | Depression | | Anorexia | |
|---|---|---|---|---|
| | Risk | Non-Risk | Risk | Non-Risk |
| 1 | hair | putt | study | item |
| 2 | weight | dispenser | **eatingdisorders** | id |
| 3 | https | tf | sex | nbsp |
| 4 | jpg | restrict_sr | sister | spoiler |
| 5 | skin | 3a | stress | men |
| 6 | bed | 27 | hair | business |
| 7 | health | keys | white | car |
| 8 | water | author | im | food |
| 9 | control | trade | **stomach** | law |
| 10 | kill | search | **unhealthy** | music |
| 11 | mother | site | 15 | win |
| 12 | youtube | data | world | state |
| 13 | baby | season | afraid | content |
| 14 | boyfriend | film | buy | message |
| 15 | knew | sex | **calorie** | fight |
| 16 | asked | movies | red | open |
| 17 | kid | sort | game | film |
| 18 | dad | books | gaining | wikipedia |
| 19 | op | children | girl | girl |
| 20 | pay | wikipedia | girlfriend | subreddit |

For each user, their posts/comments are equally divided into 10 chunks based on the chronological order. Each post/comment or WRITING includes four fields: TITLE, DATE, INFO and TEXT. TITLE is the post title. For a comment, TITLE is always empty. INFO means the source of the message. TEXT is the body of the post/comment. The number of posts/comments varies from user to user. Moreover, there is no consensus on the total time of writing. Since it is difficult to obtain the standardized time as feature, our models take the information from only TITLE and TEXT into account.

### 2.3 Evaluation

F-score and ERDE are the major metrics used in CLEF eRisk. Equation 1 shows the formula of F-score, where $\beta = 1$. ERDE complementally rewards early alerts because F-score is unaware of time. Equation 2 shows the latency cost function $lc_o(k)$, where $k$ is the number of textual items giving the answer, also called delay $k$ times, and $o$ is the parameter that controls the cost rate. The relationship between $k$ and $o$ is shown in Fig. 1. For a true negative or a true positive prediction, the ERDE is zero; for a false negative prediction, the ERDE is one; for a false positive prediction, the ERDE is set by Equation 3. In eRisk 2018, the averaged $\text{ERDE}_5$ and the averaged $\text{ERDE}_{50}$ are employed to evaluate the performance.

$$F_\beta = \frac{\left(1+\beta^2\right) \times \text{true positive}}{\left(1+\beta^2\right) \times \text{true positive} + \beta^2 \times \text{false negative} + \text{false positive}} \tag{1}$$

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

$$\text{ERDE}_{\text{true positive}} = lc_o(k) \times \text{true positive} \tag{3}$$

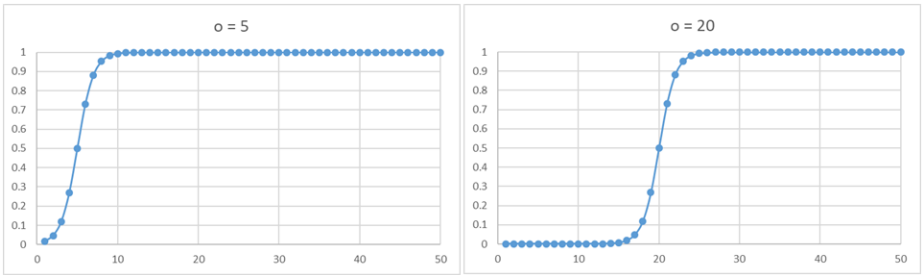



**Fig. 1.** Latency cost functions $lc_5(k)$ and $lc_{20}(k)$.

## 3 Proposed Method

We formulate the detection task as the problem of sentence classification. A classifier based on convolutional neural network (CNN) [9] is proposed and trained on the depression and the anorexia datasets. Scikit-learn [10] is also used for computing the TF-IDF for each word in both datasets.

### 3.1 Training Model

The dataflow of the training procedure is shown in Fig. 2. We first compute the TF-IDF for each word, and remove the words with low TF-IDF score in the sentence. Finally, the sentence classifier is trained with the refined sentences. The details are listed as follows.

**Keyword Selection.** See Fig. 2 (a), we select the top 300 words with the highest TF-IDF, calculated in the risk documents. The toolkit TF-IDF Vectorizer is used to index and convert each word to a unique integer in the range between 1 and 300.

**Sentence Representation.** The contents in TITLE and in TEXT from a WRITING are concatenated as a sequence of words. We discard the words other than the top 300 keywords. The rest of the sequence will be trained to encode as a vector by using the CNN-based sentence encoder. This step is important to convert an instance into a vector and an example of sentence encoding in Figure 2 (b).

**Model Training.** We regard the posts/comments written by risk users as positive instances, and those written by normal users as negative instances. Then, we train the CNN model[3] to identify the potential patients and model architecture is shown in Figure 2 (c).

[3] https://github.com/Shawn1993/cnn-text-classification-pytorch

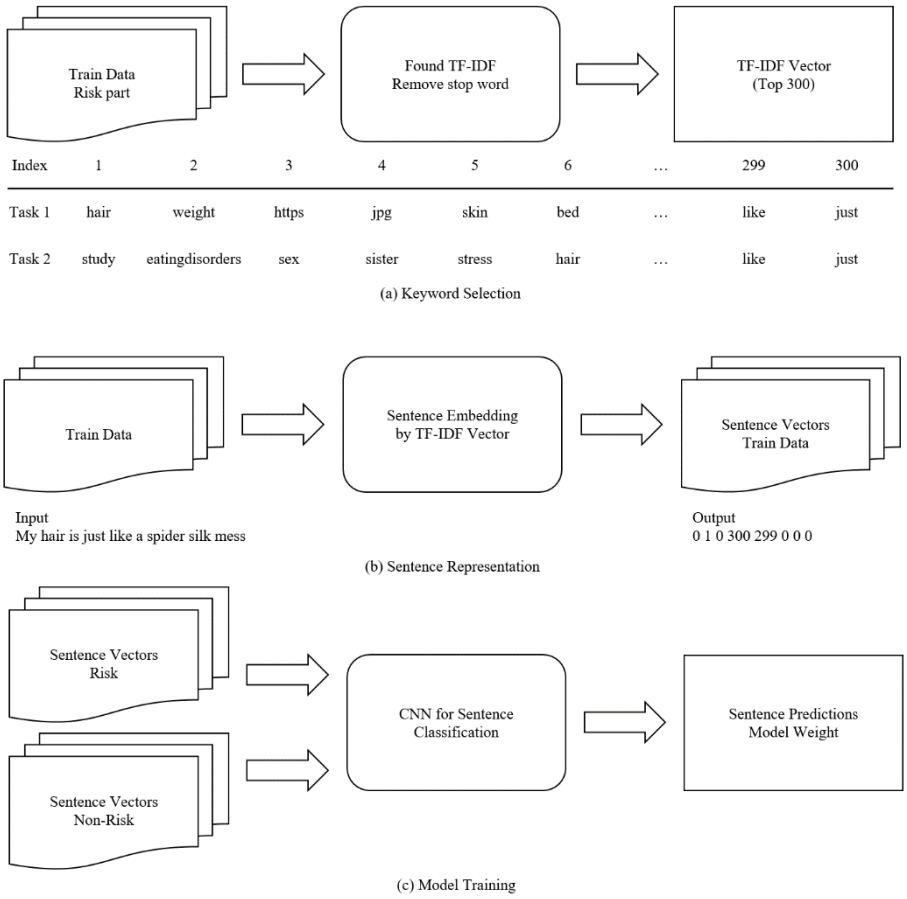
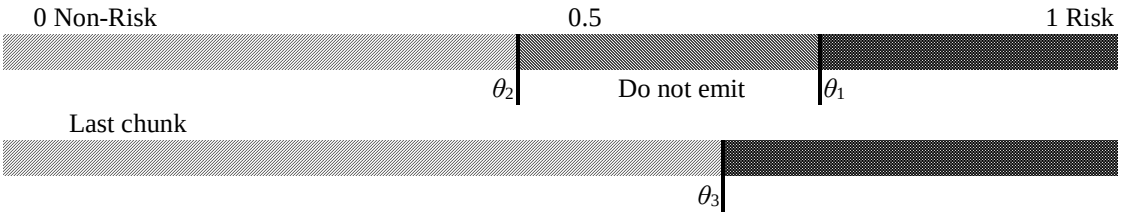



**Fig. 2.** Dataflow of the training procedure.

### 3.2 Prediction Strategy

Based on the binary classification results, we design a strategy to predict the high-risk users as early as possible. First, we perform the CNN classifier to predict every post/comment in a chunk of a user. See Table 4. We emit a risk on this user if more than $\theta_1$ of posts/comments are labeled as positive. On the other hand, we emit a non-risk on this user if less than $\theta_2$ of posts/comments are labeled as negative. Otherwise, we do not emit on this user except in the last chunk. In the last chuck, we emit a risk on the user if more than $\theta_3$ of posts/comments are labeled as positive. Otherwise, a non-risk is emitted. The thresholds $\theta_1$, $\theta_2$, and $\theta_3$ are real values between 0 and 1. We tune them with the development set.

**Table 4.** Sample of prediction threshold




## 4 Experimental Results

After the last chunk submitted, scoreboard reports [8] shows performance with $ERDE_5$, $ERDE_{50}$, Precision, Recall, and F-score. We compare our performance (denoted as TBS) with those of leading teams in the depression task and the anorexia task in Table 5 and Table 6, respectively. In terms of $ERDE_5$, the performance of our model in depression detection is better than that in anorexia detection.

There are different leading models in terms of $ERDE_5$, $ERDE_{50}$, F1, P and R. There is a tradeoff between the different goals. The model with higher F-score usually suffers from poor $ERDE_5$. In addition, the performances of the same models in the depression and the anorexia tasks are inconsistent. This result reveals the difference between these two mental illnesses. Overall, early risk detection is challenging, especially when multi-objectives are needed to optimize.

**Table 5.** Results of the depression task.

| Team | Model | $ERDE_5$ | $ERDE_{50}$ | F1 | P | R |
|---|---|---|---|---|---|---|
| UNSL | A | **8.78** | 7.39 | 0.38 | 0.48 | 0.32 |
| FHDO-BCSG | B | 9.50 | **6.44** | **0.64** | 0.64 | 0.65 |
| RKMVERI | C | 9.81 | 9.08 | 0.48 | **0.67** | 0.38 |
| TBS | A | 10.81 | 9.22 | 0.37 | 0.29 | 0.52 |
| UDC | B | 15.79 | 11.95 | 0.18 | 0.10 | **0.95** |

**Table 6.** Results of the anorexia task.

| Team | Model | $ERDE_5$ | $ERDE_{50}$ | F1 | P | R |
|---|---|---|---|---|---|---|
| UNSL | B | **11.40** | 7.82 | 0.61 | 0.75 | 0.51 |
| FHDO-BCSG | E | 11.98 | 6.61 | **0.85** | 0.87 | 0.83 |
| FHDO | D | 12.15 | **5.96** | 0.81 | 0.75 | **0.88** |
| UNSL | D | 12.93 | 9.85 | 0.79 | **0.91** | 0.71 |
| TBS | A | 13.65 | 11.14 | 0.67 | 0.60 | 0.76 |

## 5 Conclusions and Future Work

This work shows our proposed model that combines TF-IDF and CNN classification for early risk detection of depression and anorexia. In CLEF eRisk 2018, our model achieves decent $ERDE_5$ in both tasks. According to the challenging issues discussed in this paper, we will explore advanced methodologies for early risk detection. In future work, we will improve the model according to the knowledge extracted from in-domain resources such as Diagnostic and Statistical Manual of Mental Disorders (MSD-5) [11].

## References


1. Malmasi, S., Zampieri, M.: Detecting Hate Speech in Social Media: Proceedings of Recent Advances in Natural Language Processing, pages 467–472, Varna, Bulgaria, (2017)
2. Habib, M. B., Keulen, M. V.: Information Extraction for Social Media: Proceedings of Third Workshop on Semantic Web and Information Extraction, pages 9–16, Dublin, Ireland, (2014).
3. Garimella, A., Mihalcea, R.: Zooming in on Gender Differences in Social Media: Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 1–10, Osaka, Japan, (2016).
4. Samghabadi, N. S., Maharjan, S., Sprague, A., Diaz-Sprague, R., Solorio, T.: Detecting Nastiness in Social Media: Proceedings of the First Workshop on Abusive Language Online, pages 63–72, Vancouver, Canada, (2017).
5. Zirikly, A., Diab, M.: Named Entity Recognition for Arabic Social Media: Proceedings of NAACL-HLT 2015, pages 176–185, Denver, Colorado, (2015).
6. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. Proceedings Conference and Labs of the Evaluation Forum CLEF 2017, pages 346–360, Dublin, Ireland (2017)
7. Losada, D.E., Crestani, F.: A Test Collection for Research on Depression and Language Use. Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, pages 28–39, CLEF 2016, Évora, Portugal, (2016)
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. Kim, Y.: Convolutional Neural Networks for Sentence Classification: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pages 1746–1751, EMNLP 2014, Doha, Qatar, (2014)
10. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G.: API design for machine learning software: experiences from the scikit-learn project: ECML PKDD workshop: languages for data mining and machine learning, pages 108–22, (2013)
11. American Psychiatric Association: Diagnostic and statistical manual of mental disorders (5th ed), VA: American Psychiatric Association. Arlington, (2013).