Detection of early sign of self-harm on Reddit using multi-level machine

Hojjat Bagherzadeh[1], Ehsan Fazl-Ersi[2], and Abedin Vahedian[2]

[1] Dept. of Computer Engineering, Ferdowsi University of Mashhad, Iran
bagherzadehhosseinabad.hojjat@mail.um.ac.ir
[2] Dept. of Computer Engineering, Ferdowsi University of Mashhad, Iran
{fazlersi,vahedian}@um.ac.ir

Abstract. This paper describes the participation of the EFE research team in task1 of CLEF eRisk 2020 competitions. This challenge basically focuses on the early detection of symptoms of self-harm from users' posts on social media. Identifying mental illnesses especially in the early stages can help people and avoid risky behaviors. Personal notes on social media are often indicative of one's psychological state, therefore using natural language processing techniques on users' posts one can develop an early risk detection system. The proposed method is basically consisting of Word2Vec representation, an ensemble of SVM and deep neural network and also attention layers. The obtained results are very competitive and show the strength of the system provided in the early diagnosis of self-harm.

Keywords: Early Risk Detection, Self-Harm, Natural Language Processing, SVM, Attention, Word2Vec.

1 Introduction

Self-harm, also known as self-injury, is defined as intentional bodily harm and can affect all people, regardless of age, gender and, race. It can be considered as a common mental health issue, which can lead to several mental illnesses including depression, anxiety and, emotional distress [9] . Previous findings suggest that people's narratives or writing patterns can reflect their mental state [29, 30]. So with help of sentiment analysis, researchers try to identify mentally ill individuals based on people's writings on the internet and this is the main objective behind the CLEF eRisk challenges [17, 18, 20]. This year's challenge has two tasks. Task 1 deals with early detection of self-harm and task 2 tries to measure the severity of the signs of depression. The EFE team participated in Task 1 of the competition and that is given a sequence of writings for each user, the system attempts to detect signs of self-harm in users as early as possible. User's writings are processed in the same order in which they were sent and it lets chronologically monitor user's activity.

This paper presents the participation of the EFE team in self-harm early detection challenges of CLEF 2020 [22]. The method is an ensemble of deep neural networks and Support Vector Machine (SVM) as the classifier of the approach which takes its features from vector representation of the text of people posts. These vectors are Word2Vec representation of cleaned text tweaked by attention layers at different steps. Evaluation results of proposed method runs and all other competition runs of the task 1 is discussed.

The rest of the paper is organized as follows. Section 2 covers related work, while section 3 gives a brief description of Task 1 of early risk detection and the used datasets. And part 4 introduces the proposed method. Analysis of the results of experiments are presented in section 5 and finally, the conclusion and suggested directions for future works are presented in section 6.

2 Related Work

Early risk detection based on sentiment analysis is a trending research field having many growing applications. In recent years, given the popularity of social media networks as a source for news and information such health-related datasets have been made available which attracted substantial attention and led to the introduction of online competitions such as CLEF [1, 17] , CLPsych Shared Task [8, 23, 36]. Research suggests that individuals with mental issues can be identified by what they publicly share on online social media platforms because of the language patterns they use in their written texts. Thus, with advances in Natural Language Processing (NLP) researchers can now provide tools that have the capability of detecting mental illness in early stages.

NLP modules basically have two steps. The first step is a vector representation of text such as term frequency-inverse document frequency (TF-IDF) [2], pre-defined patterns, or Part-Of-Speech tagging which needs expert views over the context. Also, there are more generic text representations namely Word2Vec [25, 26], Doc2Vec [15], and LDA (Latent Dirichlet Allocation) [3, 24] that are based on counting all words of context. All of these try to represent the text as a high-dimensional vector that is appropriate for machine-learning engines. The second step is the learning process. SVM, neural networks and inference models are some of many learning models which are being wildly used in the NLP process.

Dealing with the detection of mental illness, especially self-harm is a challenging task as it is commonly relied on self-report. Most people who have self-harm also suffer from other mental illnesses, such as depression and anxiety, and thus make it difficult to be distinguished [13, 10]. Wang et al. [34] were detecting self-harm content on Flicker using word embedding and deep neural networks. UNSL team [5], one of the participants in eRisk 2019, designed a special dictionary-based text classifier. Bouarara and his team [4] analyzed users tweets to detect suicidal or self-harm behaviors to prevent any risk attempt using a sentiment classification model. Research findings state that users' writing patterns can express their mental state [7, 6, 32]. Furthermore, based on researches in this field, EFE team participated in in CLEF eRisk 2020 competition using a combination of 2 deep neural networks and SVM models trained by Word2vec text representation which promising results were obtained.

3 Dataset and Competition

The training dataset of early self-harm detection as Task 1 of CLEF 2020 was provided by the competition organizer and was the Task 2 of CLEF 2019 competition [20, 19]. The dataset consists of dated textual data of users' online posts labeled as self-harm and non-self-harm. The labeling only determines the status of each user and doesn't suggest any label for each writing. Table 3 shows a brief statistic and summary of Task 1 train data [21].

Table 1. Task 1 self-harm train dataset statistics

Train Dataset		
	Self-harm	**Control**
No. Users	41	299
Avg. No. writings per user	169	546.8
Avg. No. words per writing	24.8	18.8

The training dataset includes whole writings of each user and a label indicating the self-harm status of each one. The test stage though has an iterative strategy and a new round of writing for each user is being released only after the current run results are sent by the competitor. The evaluation measures being used in this challenge other than precision, recall, and F1, is ERDE. A detailed description of the tasks and evaluation metrics can be found in the corresponding task description paper [21].

4 Proposed Method

The proposed method for eRisk 2020 Task 1 is a multi-level approach which is a combination of deep neural networks and SVM machine along with attention mechanism, as shown in Fig. 1. At each round, first, at word level, the text cleaning steps including lowercasing, tokenizing, removing additional phrases and stemming is done to obtain a cleaned version of submitted texts and then vector representation of post using Word2Vec model [15, 25] is computed. Then, at user level these representation of posts is fed to the first level machines and scores of indicating the level of self-harness for each post is achieved. Next, at user level, these scores are aggregated to create user level features. Using an attention mechanism and Chi-Squared feature selection technique [25, 33] the most appropriate features are being selected as input to the user level learning machines. Finally, values of user level SVM machines are making the final decision based on a scoring fusion function. Therefore, at each round based on the user's writings from the beginning until this round a decision about considering the user as a self-harm case is being made. Further details for each level are as follows.

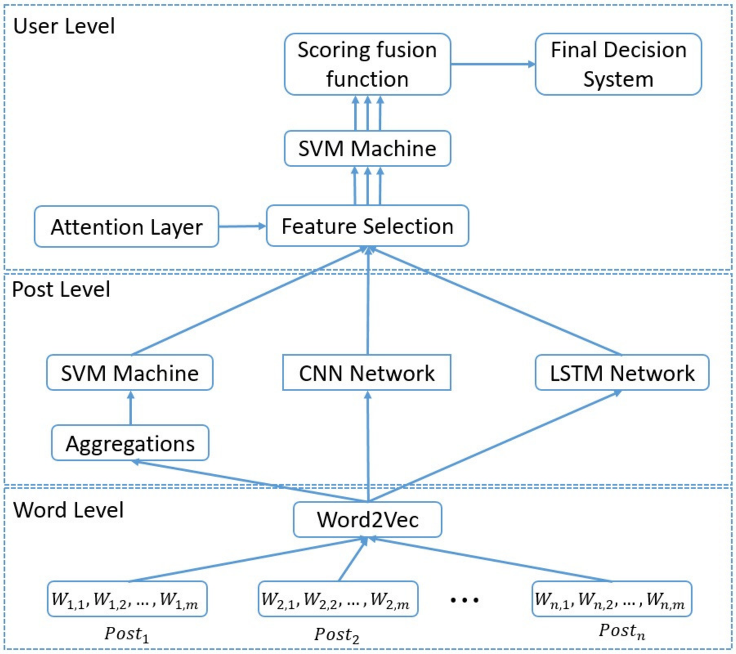

Fig. 1. Architecture of the proposed model

4.1 Word level

Word level or input layer is where create a numerical representation of words and posts used by learning machines at the next level. First, the text of people's post is tokenized and cleaned using NLP tools which mostly involve converting plurals, removing web address and hashtags, lowercasing, stemming and lemmatizing [27]. Then these clean words are fed to the word embedder which in this experiment is a customized Word2Vec model with 100-dimension vector space that is trained on Twitter and Reddit posts. Word2Vec is two-layer neural network that is trained to reconstruct or predict surrounding contexts of words in a sentence [14] and the inner layer weights of trained network are used as numeric representation of words. Therefore, each post P is represented by $[W2V_1, W2V_2, ..., W2V_m]_p$ and each $W2V_i^p$ is a 100-dimension vector which is Word2Vec representation of i-th word of p-th post.

4.2 Post level

In post level, the Convolutional Neural Network (CNN) [27, 14, 16] , Long Short-Term Memory (LSTM) [35, 12] and SVM [31, 28] are the main learning machines. The one dimensional CNN and LSTM networks process words of each post based on their chronological order and gives a score for each post which determines the probability of belonging to the self-harm class. Each $W2V_i^p$ Word2Vec representation of each words of post is fed directly to the CNN and LSTM networks. Additionally, SVM is not being able to sequentially process words' posts and thus the aggregated version of words representation which is a weighted average of each words' post as shown in equation 1 is fed to the SVM machine. Then the SVM like the other two neural machines gives a score to each post.

$$R_p = \sum_{i=1}^{n_p} (W_{word})_{p,i} \times W2V_i^p \tag{1}$$

Where R_p represents the numerical representation of p-th post and $(W_{word})_{p,i}$ is weight of i-th word of the p-th post which is computed in the training process based on the importance of every word in the degree of positivity of each post. And $W2V_i^p$ is the Word2Vec representation of i-th word of p-th post. Therefore, the output of this level for each post of a user is a three-value vector in a way that each learning machine produce one value.

4.3 User level

At user level, posts' score of each user form a $3 \times n$ matrix which n is the number of post sent by user up to the current round. In the training stage, by using common statistical measures such as average, standard deviation and variance [11] 37 features are generated for each user and among these features, with the help of Chi-squared feature selection method [25, 33] , the 7 most descriptive ones are being selected and used as inputs in the SVM machine. In the test stage, at each round, these 7 descriptive features in forms of 3×7 matrix are created based on user writings from the beginning.

Before applying statistical measures on scores of post level, scores are changed by the attention mechanisms. Considering the fact that not all the posts sent by a user is related to one's mental state, posts scores are weighted by their correlation to self-harm category. Therefore, the score's posts are being weighted and then used to generate the selected statistical features which are fed into the final SVM.

At each round, the three-digit value output of the final SVM is used by the scoring fusion function which calculates a value indicating the level of self-harm of the user based on one's writing from the beginning. These values are sent to the final decision system to alerts a '1' for users with self-harm mental status.

The final decision system works in a way that it stores scores of scoring fusion function at each round and then decide to alert '1' for users whenever the conditions depicted below are met.

- The average scores of user's up to this round.
- The number of ascending cases of user's scores.
- The number of values above the maximum threshold level.
- The average higher scores of user's up to this round.

5 Experimental setup and results

The main parts of proposed system are shown in Fig. 1. EFE team have participated in task 1 with three runs, each of which has small differences.

- The first run model is exactly as depicted in Fig. 1 and used eRisk 2018 task 1 & 2 as the training dataset for optimizing the post level and used eRisk 2019 task 1 & 2 as the training dataset for optimizing the user level and configuring the attention mechanism.
- The second run configuration is as same as the first run, and the only difference is for the training datasets. eRisk 2018 depression task was used to train the first post level of the machine and the eRisk 2019 self-harm was used to train the user level of the model.
- The third run model only has the SVM machines and the neural network models of the system are omitted. The datasets for post and user levels are the same as run 2 training sets.

Evaluation metrics for this challenge were two groups that are fully explained in [1]. The first group measures are precision, recall, and F1 metrics which consider accuracy of the model on unbalanced datasets. And the second group which is early risk detection error (ERDE), $latency_{TP}$ and $latency - weighted\ F1$ consider accuracy of the model in presence of time.

Table 5 shows the official results of the 3 runs of the proposed system. As can be seen in the table, the second run has the best performance among the others and that is because of choosing the self-harm dataset for training the user level. However, the first run has shown comparable results, which indicates the connection between depression and self-harm and other mental illnesses.

Table 2. Official result of the proposed method runs in self-harm task (T1)

Run	P	R	F1	ERDE5	ERDE50	$Latency_{TP}$	Speed	Latency-weighted F1
1	0.73	0.519	0.607	0.257	0.142	11	0.961	0.583
2	0.625	0.625	0.625	0.268	0.117	11	0.961	0.601
3	0.496	0.615	0.549	0.283	0.14	11	0.961	0.528

There are 56 runs of 12 teams participating in Task 1 of 2020 eRisk CLEF challenge. Table 3 shows the statistic of participant results compare to the proposed method. As shown in Table 5 the method has gained comparable results

Table 3. Statistics of results of 56 runs and rank of proposed method for task 1 (*: in these measures lower value means better performance.)

Run	P	R	F1	ERDE5*	ERDE50*	Latency	Speed	Latency 2	lapse time min (per writing)
Min	0.237	0.01	0.02	0.423	0.269	133	1	0.019	0.301
Max	0.913	1	0.754	0.134	0.071	1	0.526	0.658	200
Average	0.437	0.639	0.441	0.247	0.170	15.89	0.943	0.448	18
Std.	0.235	0.322	0.175	0.056	0.042	29.31	0.107	0.126	57.39
EFE Team	8	30	4	33	11	11	16	3	2

in F1 and latency-weighted F1 measure and achieved 4th rank in F1 and 3rd rand in latency-weighted F1 measure in almost the shortest time needed for processing and completing the challenge. Because of the imbalance in the dataset, the key to great performance is to maintain the balance between P and R, which is achieved in run 2 of the proposed model. Given the fact that this was the first attempt participating in such competition, we paid a lot of attention to giving early answers and as a result, the best outcomes of the model were not obtained. This also explains why the latency-weighted F1 rank is the third, while the F1 rank is the forth

6 Conclusion and future work

In this article using the presented model, EFE team participated in task 1 of eRisk2020 [21]. The task was to detect the sign of self-harm in users based on their writings as early as possible. By engaging in this challenge, the capability of social media's content as a potential source for applications related to health and safety issues has been demonstrated. The proposed system is an ensemble multi-level method based on SVM, CNN, and LSTM network, which are fine-tuned by the attention layers.

The test results show the positive effects of using attention mechanisms in the post layer and the user layer on the system, especially since not all posts sent by a person reflect his or her mental state. Another main difficulty is that there is always a trade-off between early decision making and more precise decision making. In this way, on the one hand, there is the need to detect the sign of mental illness in the user as early as possible and on the other hand, the more writings the system processes about the user, the more accurate the answer will be.

Finally, considering the fact that this is the first attempt of EFE team at such challenges, it's been found that a lot of work can be done to improve the system for real situations. Future research direction in improving the model is by working on better encoding text into numerical representation and also creating better attention mechanisms at different levels of the system. Also, another research interest is to find an optimum, under which both accuracy and giving the fastest answer are maintained.

References

1. CLEF eRisk: Early risk prediction on the Internet, `https://erisk.irlab.org/`
2. Beel, J., Langer, S., Gipp, B.: TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections (2017)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14, pp. 601–608. MIT Press (2002)
4. Bouarara, H.A.: Detection and Prevention of Twitter Users with Suicidal Self-Harm Behavior. International Journal of Knowledge-Based Organizations **10**(1), 49–61 (nov 2019)
5. Burdisso, S.G., Errecalde, M., Montes-Y-Gómez, M.: UNSL at eRisk 2019: a Unified Approach for Anorexia, Self-harm and Depression Detection in Social Media. Tech. rep. (2019)
6. Choudhury, M.D., De, S.: Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. undefined (2014)
7. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses pp. 1–10 (2015)
8. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych 2015 Shared Task: Depression and PTSD on Twitter pp. 31–39 (2015)
9. Edmondson, A.J., Brennan, C.A., House", A.O.: "non-suicidal reasons for self-harm: A systematic review of self-reported accounts". "Journal of Affective Disorders" **"191"**, "109 – 117" ("2016")
10. Gratz, K.L.: Risk factors for deliberate self-harm among female college students: The role and interaction of childhood maltreatment, emotional inexpressivity, and affect intensity/reactivity. American Journal of Orthopsychiatry **76**(2), 238–250 (apr 2006)
11. Holosko, M., Thyer, B.: Commonly Used Statistical Terms. In: Pocket Glossary for Commonly Used Research Terms, pp. 145–156. SAGE Publications, Inc. (jan 2014)
12. Jianqiang, Z., Xiaolin, G., Xuejun, Z.: Deep Convolution Neural Networks for Twitter Sentiment Analysis. IEEE Access **6**, 23253–23260 (jan 2018)
13. Kairam, S., Kaye, J., Guerra-Gomez, J.A., Shamma, D.A.: Snap decisions? How users, content, and aesthetics interact to shape photo sharing behaviors. In: Conference on Human Factors in Computing Systems - Proceedings. pp. 113–124. Association for Computing Machinery (may 2016)
14. Kshirsagar, R., Morris, R., Bowman, S.: Detecting and Explaining Crisis (2017)
15. Le, Q.V., Mikolov, T.: Distributed Representations of Sentences and Documents. 31st International Conference on Machine Learning, ICML 2014 **4**, 2931–2939 (may 2014)

16. Liao, S., Wang, J., Yu, R., Sato, K., Cheng, Z.: CNN for situations understanding based on sentiment analysis of twitter data. In: Procedia Computer Science. vol. 111, pp. 376–381. Elsevier B.V. (jan 2017)
17. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: Experimental foundations. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 10456 LNCS, pp. 346–360. Springer Verlag (2017)
18. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations. In: Jones, G.J.F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 346–360. Springer International Publishing, Cham (2017)
19. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 Early Risk Prediction on the Internet. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 11696 LNCS, pp. 340–357. Springer Verlag (sep 2019)
20. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk at CLEF 2019: Early Risk Prediction on the Internet. In: Linda Cappellato Nicola Ferro, D.E.L.H.M. (ed.) Conference and Labs of the Evaluation Forum. CEUR-WS.org (2019)
21. Losada, D.E., Crestani, F., Parapar, J.: eRisk 2020: Self-harm and depression challenges. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 12036 LNCS, pp. 557–563. Springer (apr 2020)
22. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2020: Early Risk Prediction on the Internet. In: A. Arampatzis E. Kanoulas, T.T.S.V.H.J.C.L.C.E.A.N.L.C.N.F.e. (ed.) Experimental Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association. Springer International Publishing (2020)
23. Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., Schwartz, H.: CLPsych 2018 Shared Task: Predicting Current and Future Psychological Health from Childhood Essays. pp. 37–46 (2018)
24. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning Word Vectors for Sentiment Analysis. Tech. rep. (2011)
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems (oct 2013)
26. Mikolov, T., Yih, W.T., Zweig, G.: Linguistic Regularities in Continuous Space Word Representations. Tech. rep. (2013)
27. Mohan, V.: Preprocessing Techniques for Text Mining - An Overview (feb 2015)
28. Monika, R., Deivalakshmi, S., Janet, B.: Sentiment Analysis of US Airlines Tweets Using LSTM/RNN. In: Proceedings of the 2019 IEEE 9th International Conference on Advanced Computing, IACC 2019. pp. 92–95. Institute of Electrical and Electronics Engineers Inc. (dec 2019)
29. Moulahi, B., Azé, J., Bringay, S.: Dare to care: A context-aware framework to track suicidal ideation on social media. pp. 346–353 (10 2017)
30. Paul, M., Dredze, M.: You Are What Your Tweet: Analyzing Twitter for Public Health. Artificial Intelligence **38**, 265–272 (01 2011)
31. Ragheb, W., Azé, J., Bringay, S., Servajean, M.: Attentive Multi-stage Learning for Early Risk Detection of Signs of Anorexia and Self-harm on Social Media. Tech. rep. (2019)

32. Schwartz, H., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.: Towards Assessing Changes in Degree of Depression through Facebook (jan 2014)
33. Sun, J., Zhang, X., Liao, D., Chang, V.: Efficient method for feature selection in text classification. In: Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017. vol. 2018-Janua, pp. 1–6. Institute of Electrical and Electronics Engineers Inc. (mar 2018)
34. Wang, Y., Tang, J., Li, J., Li, B., Wan, Y., Mellina, C., O'Hare, N., Chang, Y.: Understanding and Discovering Deliberate Self-Harm Content in Social Media. In: Proceedings of the 26th International Conference on World Wide Web. pp. 93–102. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017)
35. Wang, Y., Zhang, C., Zhao, B., Xi, X., Geng, L., Cui, C.: Sentiment Analysis of Twitter Data Based on CNN. Shuju Caiji Yu Chuli/Journal of Data Acquisition and Processing **33**(5), 921–927 (sep 2018)
36. Zirikly, A., Resnik, P., Uzuner, .O., Hollingshead, K.: CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. Tech. rep. (2019)