USDB at eRisk 2020: Deep learning models to measure the Severity of the Signs of Depression using Reddit Posts

Amina MADANI[1], Fatima BOUMAHDI[1], Anfel BOUKENAOUI[1], Mohamed Chaouki KRITLI[1], and Hamza HENTABLI[2]

[1] LRDSI Laboratory, BLIDA 1 University, Blida , Algeria
{a_madani,f_boumahdi}@esi.dz, anfelboukenaoui@gmail.com, kritlichawki56@gmail.com
[2] Information Assurance and Security Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
hentabli_hamza@yahoo.fr

Abstract. In this paper, we describe the participation of our USDB group (University of Saad Dahleb Blida) in the shared task T2 of the eRisk Lab at the CLEF 2020 workshop. This task focused on measuring the severity of the signs of depression from a thread of user posts. In response to this task, we study the performance of two different deep learning models (CNN and BiLSTM) in order to provide more perspectives for depression researches.

Keywords: Depression severity · Social networks · Natural language processing · Deep learning · CNN · BiLSTM · Sentiment analysis.

1 Introduction

Depression identification has been the subject of research of many fields, psychiatry, psychology, medicine and even sociolinguistics fields. Depression comes in different degrees and the examinations are usually done through one of the popular questionnaires used by psychologists, such as the Center of Epidemiologic Studies Depression Scale (CES-D) [26], Beck's Depression Inventory (BDI) [4] and Zung's Self-Rating Depression Scale (SDS) [38]. But, these examinations lack empirical data as they use the patient's observations or a third-party's ones which puts the results under the risk of flawed subjective human testing that can be manipulated easily, often with the purpose of gaining antidepressants or just to hide one's own depression from peers [21].

Twitter, Facebook and Reddit are different social media platforms that allow people to share their opinions and their personal thoughts. It has been proven

that such data can be used to study clinical matters, especially when it comes to mental illnesses like depression.

Many research works have analyzed the prowess of these data for determining indications of depression. Furthermore, the scientifc community has set forth different shared tasks like eRisk (Early Risk Prediction on the Internet) of CLEF (Conference and Labs of the Evaluation Forum).

In this paper, we describe the participation of our team USDB to the CLEF eRisk 2020 task 2. The goal of the task was to measure the severity of the signs of depression, considering a set of user posts on Reddit. Participants had to answer the standard BDI depression questionnaires using the text of postings for 70 users.

Our team proposed an approach based on deep learning models to automatically fill the BDI questionnaire that are Convolutional Neural Networks model (CNN) [15] and Bidirectional Long Short Term Memory model (BiLSTM) [3,35,36], which is a type of recurrent neural network. Subsequently, to generate different runs, we use two statistical methods. For the first time, neural networks are used for the task of measuring the severity of the signs of depression.

The rest of this paper is organized as follows: section 2 presents related works that tackled the same problem as ours. In section 3, we explore our proposed approach. The following three sections are dedicated to the description of the dataset, the metrics evaluation used and the results obtained. Finally, we conclude the paper and discuss future perspectives for our proposed approach.

2 Related work

Since 2017 until now a shared task on eRisk has been organised. In 2017 and 2018, the challenge consists in performing a task on early risk detection of depression. Several researchers were focused on detecting depression. [9,28,30,19,22,27,33,31,7] are different approaches that propose interesting models and evaluate them using eRisk dataset.

In 2019, a new task was added. Measuring the severity of the signs of depression consists of estimating the level of depression from a thread of user submissions. The results of all participating teams can be found in [17]. [29,2,6,32] are several works of different teams.

In paper [29], authors proposed a rule-based method that combines machine learning with psycholinguistics and behavioural patterns. They divided the 21 questions into 6 groups that are : Depression, Guilt, Appetite, Anxiety, Fatigue and Sleep. Based on the presence of occurrences of the features considered for each group, they produced responses for each user.

In their work, [2] extracted features from users using the GPT-1 (Generative Pre-trained Transformer version 1) language model [25,37] and Linguistic Inquiry Word Count tool (LIWC) [23]. They predict responses in two ways, an unsupervised and supervised one. For the unsupervised manner, they used an approach based on vectorial representation of the user and vectorial representation of possible response using GPT-1. Cosine Similarity between vectors was calculated

to choose the user's response. For the supervised way, they used data of a PhD study, where Psychology students answer the BDI and other questionnaires and also completed parts of writing about a negative personal problem. Next, they trained support vector machines using the training data. They also submitted another supervised approach that used GPT-1 features and AutoSklearn [10]. Paper [2] concluded that without training dataset, tasks were not easy and are unsupervised and data must be annotated to improve the quality of depression prediction.

The Paper of [6] used SS3 [5] which is a word-based classifier that estimates risk based on term statistics. To train SS3, they adapted the dataset of eRisk 2018 depression detection task. They transformed the output of SS3 from a 2-dimensional vector into a BDI depression rank. All the questions were answered with 0 for persons whose depression rank was less or equal to 0. For the others, they used different methods (based on textual hint, word matching...). SS3 obtained the best AHR and ACR values, and the second-best ADODL and DCHR.

To automatically fill in the BDI questionnaire, the authors of [32] developed four models. They submitted only the results of the fourth model. The models are:

- Word polarity model using the Multi Perspective Question Answering (MPQA) subjectivity lexicon [34,1].
- Mutual information model by creating a training dataset from Reddit and using the mutual information measure to extract important tokens from depressive messages [14].
- Semantic similarity model which is based on post-level representation. The pre-trained GloVe word embeddings [24] is used to represent the words.
- In the fourth model, the results of the three models are combined using voting.

[16] say that no team was able to reach best results for each of the evaluation measures because of the difficulty of the challenge and probably the similarity of the approaches.

3 Method

In this section, we will introduce the architecture of our proposed approach (see Fig. 1). First, we do preprocessing of posts. We extract keywords that contain most important information by removing special characters, punctuations, URL and stop-words. Words would all be stemmed and lemmatized to remove noise from posts.

Some of the publications are longer or shorter. Padding is then necessary, because we need to have the inputs with the same size. We fixed the sequence length of posts to 250 and shorter input sequences are padded with zeros.

Next, we transform distributed representations of words in a vector space using the Skip-gram model [20] which is used to predict the context word for a

given target word. For a given sequence of words, the objective is to find word representations that are useful for predicting the surrounding words in a post.

After that, the sentences are encoded by means of a CNN or a Bi-LSTM model. Our first method is based on a CNN model, one of the most popular deep neural networks. Our model consists of :

- two one-dimensional convolutional layers,
- max-pooling layer : maintains only the most important words in each feature obtained from the previous layers,
- two others one-dimensional convolutional layers,
- fully connected input layer: flattens the outputs of the convolutional layer to map them into a single vector,
- fully connected layer:applies weights to predict the correct label,
- fully connected output layer:gives the final probabilities.

In these layers, each linear activation is run through ReLU (Rectified Linear Unit). The rectified linear activation function will output the input directly if is positive, otherwise, it will output zero.

Our second method involves the usage of Recurrent Neural Networks (RNN) [8,12] with Long BiLSTM [11] to make predictions on sequences of texts. LSTM [13] is used in different problems due to their ability to remember information over long periods of time. LSTM use 3 gates to capture Long Term Dependencies: *Forget Gate* for adding information to the cell state, *Input Gate* to choose what component of previous cell state must be forgotten and *Output Gate* to ascertain information to output at current cell state. The BiLSTM model brings the advantage of maintaining two separate states for inputs using two different LSTMs. The first LSTM is going forward from the beginning of the sentence, while in the second LSTM, the input sequence are fed in backward. BiLSTM allows capturing information of surrounding inputs and learning faster than LSTM model.

Last, for the two deep learning models we have a final layer, in which the representation is fed to the final fully connected softmax layer as an output feature vector. The results will be decimal probabilities for each answer. Each publication has only one answer for each question.

For each post, our two models generate 21 outputs that are answers to the BDI's questions. Finally, in order to know the level of depression for a user, we use two statistical methods to generate 4 runs. The two first runs concern the CNN model when the BiLSTM model is applied for the others. For run1 (CNN_max) and run3 (BiLSTM_max), we calculate for a question the frequency of each generated answer to choose the most recurrent answer, which makes it as relevant as possible. For run2 (CNN_suite) and run4 (BiLSTM_suite), we calculate for a question, the higher length sequence of a same answer for all generated answers from which we choose the answer of the higher value to be a response of this question.

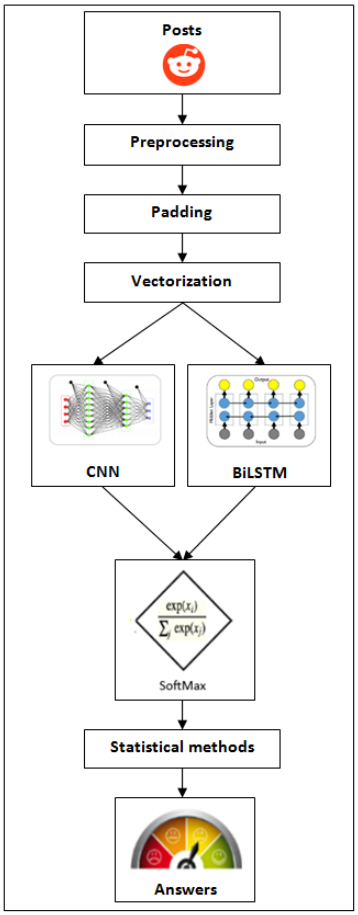

Fig. 1. The architecture of our proposed approach.

4 Dataset description

eRisk 2020 Task 2 is a continuation of 2019's task 3. This year, a dataset with 70 files was provided. Each file contains a set of posts of one user on Reddit. In Table 1, we present the characteristics of the dataset.

Table 1. Statics on the eRisk 2020 Task 2 dataset.

Number of users	70
Number of posts	35562
Average number of posts per user	508
Min number of posts per user	25
Max number of posts per user	1355

Based on a user's history of posts, task 2 was aimed to estimate the depression level of a user by automatically answering each individual question derived from the BDI questionnaire. The questionnaire has 21 questions in different classes of feelings like sadness, pessimism, crying, loss of energy, etc.

The possible responses are 0, 1a, 1b, 2a, 2b, 3a, 3b for questions 16 and 18 and 0, 1, 2, 3 for the rest of the questions.

The 2019's questionnaires and their golden truth responses were provided. Thus, they can be used as a training dataset.

5 Metrics evaluation

Four metrics are used for results evaluation:

- **Average Hit Rate (AHR)**: where HR calculates the percentage of cases where our answers are the same as the golden truth responses.
- **Average Closeness Rate (ACR)**: computes the absolute difference between our answer and the true one.
- **Average Difference between Overall Depression Levels (ADODL)**: for a user, the absolute difference between the generated overall depression score (sum of all the answers) and the real score is calculated.
- **Depression Category Hit Rate (DCHR)**: Four depression categories based on the sum of all answers of the 21 questions can be found. The categories are minimal depression, mild depression, moderate depression and severe depression. DCHR verify the depression category of the real questionnaire with the category of the automated obtained answers.

More details about these used measures and examples are given in [17].

6 Results

In this year, 5 teams submitted 17 different runs to the eRisk task 2. Our team submitted 4 runs to this task. Table 2 shows our results comparing to the best results for each metric. The results of all participants can be found in [18]. We observe that no single team was able to achieve the best results for each of the four metrics evaluation.

Comparing only our runs, run 2 and run 4 did not perform well on the dataset. Although, run 1 performed the best in the AHR with **34.97%** of the answers right and the best in the DCHR which is able to predict the correct depression severity category for **25.71%** of the users. In contrast to the 3 runs, run 3 had higher ACR (**67.78%**) and ADODL (**79.30%**) scores. Therefore, we notice that using the first statistical method based on the frequency of answers is better than using the second method based on the higher length sequence of answers.

Most importantly, we believe combining the CNN model with the BiLSTM one could improve the feature extraction process and enhance the model's performance to predict better results.

Table 2. Evaluation of our runs along with the best results achieved in task 2.

	AHR	ACR	ADODL	DCHR
Run1:CNN_max	**34.97%**	67.19%	76.85%	**25.71%**
Run2:CNN_suite	32.79%	66.08%	76.33%	17.14%
Run3:BiLSTM_max	34.01%	**67.78%**	**79.30%**	22.86%
Run4:BiLSTM_suite	33.54%	67.26%	78.91%	20.00%
Best scores	*38.30%*	*69.41%*	*83.15%*	*35.71%*

7 Conclusion

The aim of this article is to exploit artificial intelligence's deep learning models in order to measure automatically the severity of the signs of depression from an individual's posts. We described the participation of our research group at task 2 of the CLEF eRisk 2020 using using the CNN model, the BiLSTM model and statistical methods to generate runs.

We conclude that no run was able to predict overall depression better than other because this task is not easy and a training dataset with 20 users is not sufficient.

For future work, we plan to principally combine the CNN model with the BiLSTM one and we will analyze in more details the obtained results.

References

1. Mpqa resources. `http://mpqa.cs.pitt.edu/#subj_lexicon`

2. Abed-Esfahani, P., Howard, D., Maslej, M., Patel, S., Mann, V., Goegan, S., French, L.: Transfer learning for depression: Early detection and severity prediction from social media postings. In: CLEF (Working Notes) (2019)
3. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). pp. 747–754 (2017)
4. Beck, A., Ward, C., Mendelson, M., Mock, J., Erbaugh, J.: An inventory for measuring depression. archives of general psychiatry, vol. 4 (1961)
5. Burdisso, S.G., Errecalde, M., Montes-y Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications **133**, 182–197 (2019)
6. 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. In: CLEF (Working Notes) (2019)
7. Cacheda, F., Iglesias, D.F., Nóvoa, F.J., Carneiro, V.: Analysis and experiments on early detection of depression. CLEF (Working Notes) **2125** (2018)
8. Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–211 (1990)
9. Fatima, B., Amina, M., Nachida, R., Hamza, H.: A mixed deep learning based model to early detection of depression. Journal of Web Engineering pp. 429–456 (2020)
10. Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F.: Efficient and robust automated machine learning. In: Advances in neural information processing systems. pp. 2962–2970 (2015)
11. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks **18**(5-6), 602–610 (2005)
12. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in neural information processing systems. pp. 545–552 (2009)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Kraskov, A.S., Stogbauer, H.: H. & grassberger. Estimating mutual information. Phys. Rev. E **69**(6) (2004)
15. LeCun, Y., Bengio, Y., et al.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks **3361**(10), 1995 (1995)
16. Losada, D.E., Crestani, F., Parapar, J.: Early detection of risks on the internet: an exploratory campaign. In: European Conference on Information Retrieval. pp. 259–266. Springer (2019)
17. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2019 early risk prediction on the internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 340–357. Springer (2019)
18. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk 2019 early risk prediction on the internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer (2020)
19. Maupomé, D., Meurs, M.J.: Using topic extraction on social media content for the early detection of depression. CLEF (Working Notes) **2125** (2018)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

21. Nadeem, M.: Identifying depression on twitter. arXiv preprint arXiv:1607.07384 (2016)
22. Paul, S., Jandhyala, S.K., Basu, T.: Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In: CLEF (Working Notes) (2018)
23. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., Booth, R.J.: The development and psychometric properties of liwc2007: Liwc. net. Google Scholar (2007)
24. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
25. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
26. Radloff, L.S.: The ces-d scale: A self-report depression scale for research in the general population. Applied psychological measurement **1**(3), 385–401 (1977)
27. Ragheb, W., Moulahi, B., Azé, J., Bringay, S., Servajean, M.: Temporal mood variation: at the clef erisk-2018 tasks for early risk detection on the internet (2018)
28. Stankevich, M., Isakov, V., Devyatkin, D., Smirnov, I.: Feature engineering for depression detection in social media. In: ICPRAM. pp. 426–431 (2018)
29. Trifan, A., Oliveira, J.L.: Bioinfo@ uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In: CLEF (Working Notes) (2019)
30. Trotzek, M., Koitka, S., Friedrich, C.M.: Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering (2018)
31. Trotzek, M., Koitka, S., Friedrich, C.M.: Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In: CLEF (Working Notes) (2018)
32. Van Rijen, P., Teodoro, D., Naderi, N., Mottin, L., Knafou, J., Jeffryes, M., Ruch, P.: A data-driven approach for measuring the severity of the signs of depression using reddit posts. In: CLEF (Working Notes) (2019)
33. Wang, Y.T., Huang, H.H., Chen, H.H.: A neural network approach to early risk detection of depression and anorexia on social media text. In: CLEF (Working Notes) (2018)
34. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of human language technology conference and conference on empirical methods in natural language processing. pp. 347–354 (2005)
35. Zhang, Y., Wang, J., Zhang, X.: Ynu-hpcc at semeval-2018 task 1: Bilstm with attention based sentiment analysis for affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 273–278 (2018)
36. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 207–212 (2016)
37. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision. pp. 19–27 (2015)
38. Zung, W.W., Richards, C.B., Short, M.J.: Self-rating depression scale in an outpatient clinic: further validation of the sds. Archives of general psychiatry **13**(6), 508–515 (1965)