A RoBERTa-based model on measuring the severity of the signs of depression

Shih-Hung Wu[1] and Zhao-Jun Qiu[1]

[1] *Chaoyang University of Technology, Taichung, Taiwan (R.O.C)*

Abstract

In this paper, we describe our approach to the CLEF 2021 lab eRisk Task 3: Measuring the severity of the sign of depression. The main purpose of this task is to automatically measure the severity of the user's depression by analyzing the user's posting on social media. We adopt the deep learning pretrained language model, RoBERTa, as the basis of our system and propose two different approaches as the post-processing and submit 3 runs. The two post-processing weighting mechanisms is designed to make the system that will give prediction on higher level of severity. This is according to our observation on the results of last year eRisk lab that systems tend to give lower level of severity. With a fixed weighting approach, our second run gives the best Average Difference between Overall Depressions Levels (ADODL) and Depression Category Hit Rate (DCHR) this year.

Keywords

Deep Learning, RoBERTa

1. Introduction

Social media is popular, it can be seen that with the spread of mobile networks, people use social media more frequently. According to DIGITAL 2021: GLOBAL OVERVIEW REPORT [1], social media users have reached more than half of the global population. People express emotions through social media has become a daily habit.

Researcher can analyze these postings with natural language processing technology and get useful results. In eRisk Task 3: Measuring the severity of the sign of the sign of depression, systems try to predict the severity of a user's depressive symptoms by analyzing the user's postings on social media. Similar studies have been conduct on other social media, such as Facebook language predicts depression in medical records [2] and forecasting the onset and course of mental illness with Twitter data [3], which have shown the importance of evaluating user depression levels through social media postings.

The main mission of eRisk 2021 Task 3 is to explore the feasibility of automatically estimating the severity of multiple symptoms associated with depressive symptoms. The organizers estimate a user's level of depression by the user's response to each question in the questionnaire of Baker's Depression List (BDI), which assesses the existence of feelings such as sadness, pessimism, and lack of energy. The questionnaire has 21 questions, each with four answers (from 0 to 3) or seven answers (0,1a, 1b, 2a, 2b, 3a, 3b). The system performance will be assessed by the overlap between the questionnaire filled out by real users and the questionnaire filled out by the system (number of correct predictions) [4].

This is the third time that the task of depression prediction is held in eRisk lab. In the past eRisk Tasks on depression prediction, many teams have come up with different ways to study this topic, such as, the USDB team used two different deep learning models (CNN and BiLSTM) [12], the iLab team focused on the pre-processing aspects of training data [13], the RELAI team used topic model (LDA

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: shwu@cyut.edu.tw (A. 1); s10827617@gm.cyut.edu.tw (A. 2)
ORCID: 0000-0002-1769-0613 (A. 1); 0000-0002-4616-9624 (A. 2)



CEUR Workshop Proceedings (CEUR-WS.org)

and Anchor) [14] to conduct the research, the BioInfo@UAVR teams used the classifier of Yates et al. [15] and they had trained before to predict whether users were depressed [16].

Most of the previous works focus on training using different models, or pre-processed data. Our approach this time, mainly focus on post-processing, after we used state-of-the-art deep learning pretrained language model, RoBERTa, as the basis of our system. We propose two different approaches as the post-processing and submit 3 runs. According to our observation on the results of last year eRisk lab that systems tend to give lower level of severity. Our post-processing weighting mechanisms is designed to make the system that will give prediction on higher level of severity.

The rest of this article is organized as follows: Section 2 describes how eRisk Task 3 provides data and how to evaluate system. The methodology is described in Section 3, which reports our research process and our experimental settings. The last two sections explore results we have come up with, as well as the future direction of the study.

2. Data and Observation

The organizers of 2021 eRisk T3 provide the test dataset of 2019 and 2020, and as the training data. The 2020 dataset has a total of 20 users, during system developing phrase, we use the 2020 data as the training set to train the model and the 2019 dataset as the validation set to test the model. The dataset includes a severity questionnaire assessment of depressive symptoms, as well as postings on social media of a user's daily life. The questionnaire consists of a total of 21 questions and has four answers for each question, excepting that questions 16 and 18 has 7 answers [4].

eRisk T3 uses four different scoring metrics to evaluate the model, namely Average Hit Rate (AHR), Average Closeness Rate (ACR), Difference between Overall Depressions Levels (DODL) Average (ADODL) and Depression Category Hit Rate (DCHR) [8]. During system development, we focus on the AHR and DCHR metrics, since the other two metrics are relative metrics of these two metrics. We believe that optimize the two metrics will also optimize the other two.

- Average Hit Rate (AHR): For each user in the 21 questions, if the system predicted the actual result of the user in ten questions, the hit rate is 10/21, and AHR is the average hit rate to all users.
- Depression Category Hit Rate (DCHR): System predicts the questionnaire results obtained an estimate of recognized depression, which matches the assessment information obtained in the actual questionnaire.

Run	AHR	ACR	ADODL	DCHR
BioInfo@UAVR	**38.30%**	69.21%	76.01%	30.00%
iLab run1	36.73%	68.68%	81.07%	27.14%
iLab run2	37.07%	**69.41%**	81.70%	27.14%
iLab run3	35.99%	69.14%	82.93%	34.29%
prhlt_logreg_features	34.01%	67.07%	80.05%	**35.71%**
prhlt_svm_use	36.94%	69.02%	81.72%	31.43%
prhlt_svm_features	34.56%	67.44%	80.63%	**35.71%**
svm_features	34.56%	67.44%	80.63%	**35.71%**
relai_context_paral_user	36.80%	68.37%	80.84%	22.86%
relai_context_sim_answer	21.16%	55.40%	73.76%	27.14%
relai_lda_answer	28.50%	60.79%	79.07%	30.00%
relai_lda_user	36.39%	68.32%	**83.15%**	34.29%
relai_sylo_user	37.28%	68.37%	80.70%	20.00%
Run1_resultat_CNN_Methode_max	34.97%	67.19%	76.85%	25.71%
Run2_resultat_CNN_Methode_suite	32.79%	66.08%	76.33%	17.14%
Run3_resultat_BILSTM_Methode_max	34.01%	67.78%	79.30%	22.86%
Run4_resultat_BILSTM_Methode_suit	33.54%	67.26%	78.91%	20.00%
all 0s	36.26%	64.22%	64.22%	14.29%
all 1s	29.18%	73.38%	81.95%	25.71%
random (avg 1000 repetitions)	23.94%	58.44%	75.22%	26.53%

Figure 1: CLEF 2020 eRisk: Task2. Performance Results [5]

According to the 2020 CLEF eRisk results in Fig. 1 [5], the all 0' and all 1's prediction results were 36% and 29% in AHR, respectively. However the percentage dropped to 14% and 25% when evaluating DCHR, from which we speculate that the actual forecast data is tent to a higher level of serenity. That is, training set shows that users will give answers to each question with a lower level of serenity, but the overall serenity is not that low.

Table 1 shows the statistics of the training data, each user's postings is labelled according to the user's answers in the questionnaire, and the chart shows that most of the statistics are slightly biased to lower level of severity, so we expect to weight the results during the post-processing process to make the results more prone to higher level of severity will give better overall result.

Table 1

(a) The percentage of the posting distribution for the 2020 dataset, each posting is labelled based on the results of the user's answer to the questionnaire. Assuming that the number of postings of the users who answer 0 to question 1 of the questionnaire is 350, and the total number of postings all users is 1000, the percentage is equal to (350/1000 = 35%)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	19	20	21
0	35	26	24	**40**	**43**	**62**	31	29	**57**	**45**	**38**	**39**	**43**	**46**	20	38	**35**	27	**71**
1	**51**	**44**	**50**	29	38	22	20	**35**	34	32	36	32	27	27	**43**	**39**	28	**34**	22
2	9	22	20	29	14	10	**35**	32	6	7	17	11	20	23	27	15	25	26	1
3	6	8	5	8	5	6	14	3	3	15	9	18	10	4	10	9	13	13	6

(b) The percentage of the posting distribution on Question 16 and 18

	16	18
0	0	0
1a	11	**37**
1b	**49**	22
2a	9	12
2b	16	14
3a	3	6
3b	13	8

3. System Architecture

Fig. 2 shows our system flowchart. As mentioned in previous sections, we use a pre-trained model to give the Run1 and weighting the Run1 results into other two runs. The Preprocessing is quite simple, our system just delete URL, special characters, and white space from the users' postings. And sent it to BERT or RoBERTa model.

We build one model for each question, therefore, there are 21 models. Each posting is labelled with the answer of the user to the question. This labeling is assuming that each posting will give the same information on the choice of the user. We train the classifier by BERT/RoBERTa models according to the sentences in the training set. For each question, we train one classifier to decide whether one sentence lead to which answer. Since each author writes a lot of sentences, our system aggregate the vote of each sentence as our system output. In the first run, our system works with a majority vote principle, one answer will be selected if it get most votes. In the second run, we emphasize the weight of the votes by weighting more on the answers with serious results. That is, tend to be more depression. The weights are 1 to 7 for the votes of 0 to 6 respectively. In the third run, we further lower the weight of vote to 0 by rules.

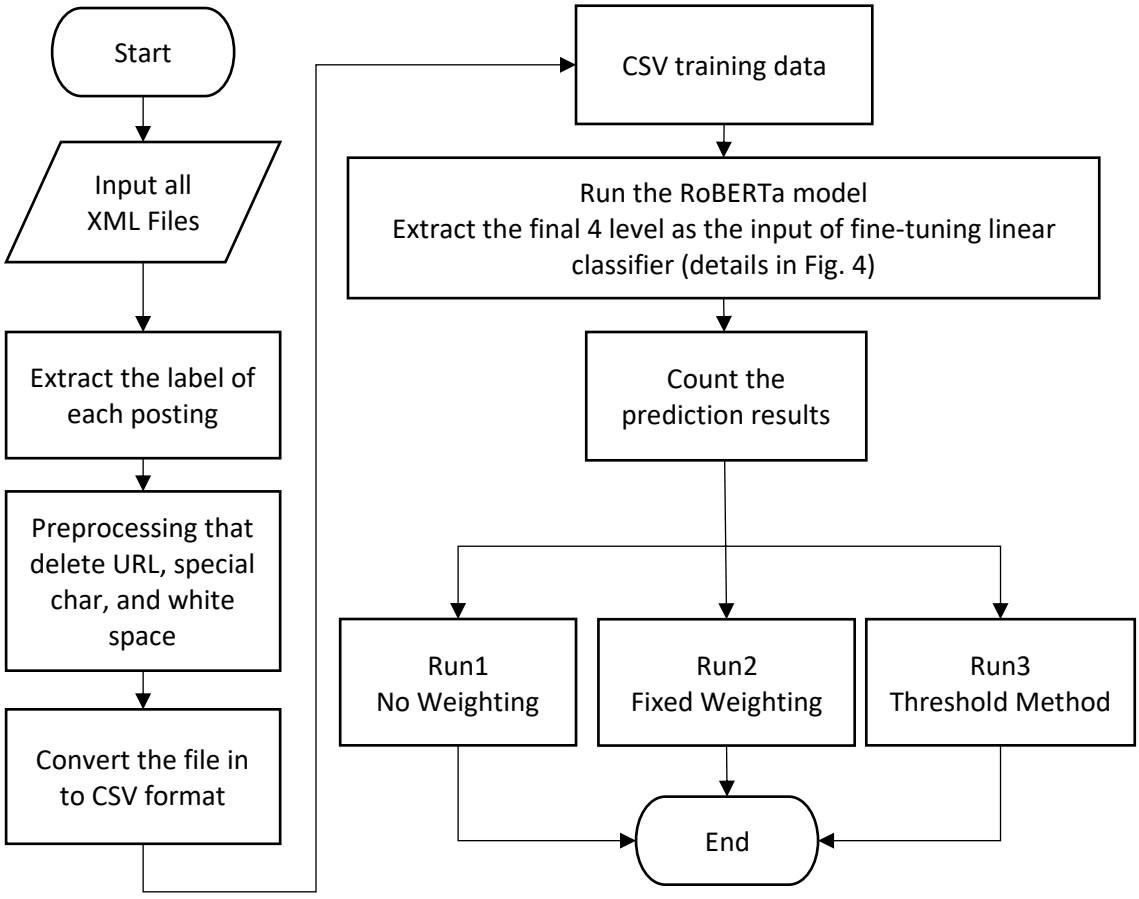


Figure 2: Our System flowchart

3.1. Data Processing

The training data contains data from 70 volunteers with questionnaire results as well as posts or comments on their social networks. The XML file format is shown in Fig. 3. We extract the TEXT content into a CSV file as our training data. We remove the URL, path, special characters, and each of the comments is organized into one line, saved in the first column. We aggregate all the posts from each anonymous ID and associate them to the user's questionnaire results, the answer to the question in order after the first column. We get a total of 33,155 comments, we use 80% (26524) for training, 20% (6631) for verification during our system development phrase.

```
<INDIVIDUAL>
<ID> ... </ID>
<WRITING>
<TITLE> ...   </TITLE>
<DATE> ... </DATE>
<INFO> ... </INFO>
<TEXT> ...  </TEXT>
</WRITING>
<WRITING>
....
</INDIVIDUAL>
```

Figure 3: The XML format of each post in the dataset[4], where ID is the anonymous user ID, TITLE is the post title, INFO is the source, and TEXT is the content of the post

3.2. Pre-trained Model

Since BERT gives good results in several natural language processing applications in recent years [17], so we adopt the BERT pre-trained model as our basis of our system. At first, we chose the pre-trained model "bert-base-uncased" and made improvements by referring to the methods of [6] and [7]. Instead of using the final output of the BERT model directly, our system extracts the output of the last four hidden layers as the input vector for linear classification, as shown in Fig. 4. The Hyper-parameters of our model is: Hidden size=768, Learning r=1e-5, weight_decay=1e-2, Epoch=5

Fig. 5 Shows the result of the model on Q3-Q6, since more epoch do not give better result, we limited our fine-tuning epoch to 5.

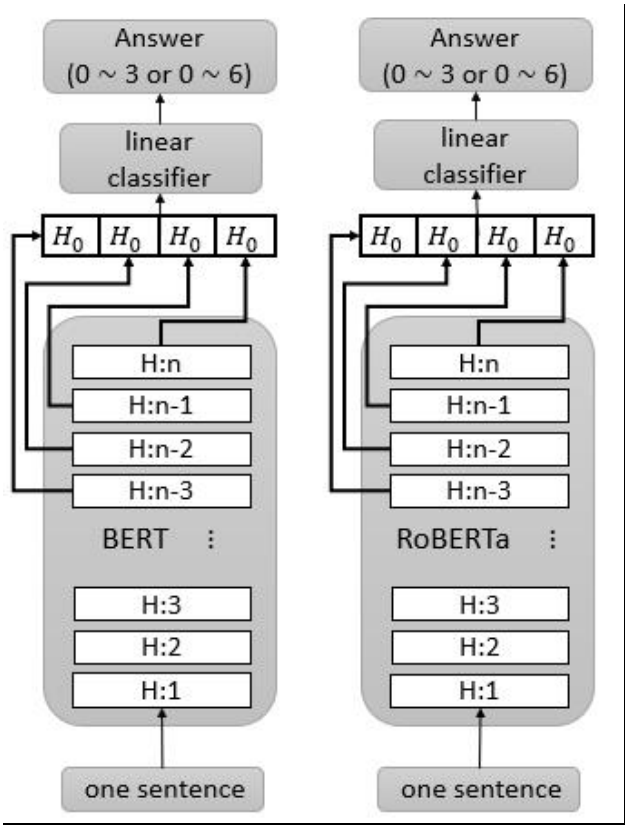


Figure 4: Our system extracts the output vector from the last four layers of the model's hidden layer and joins the four output vectors as the input vector of the linear classifier

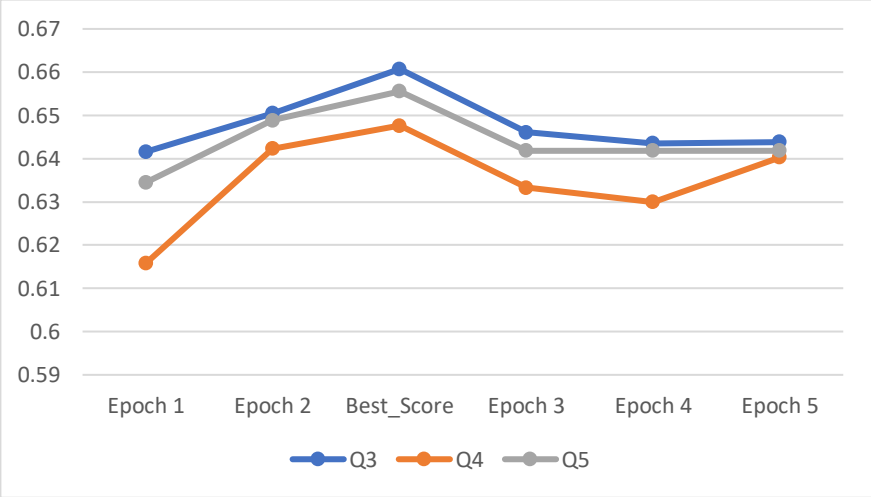
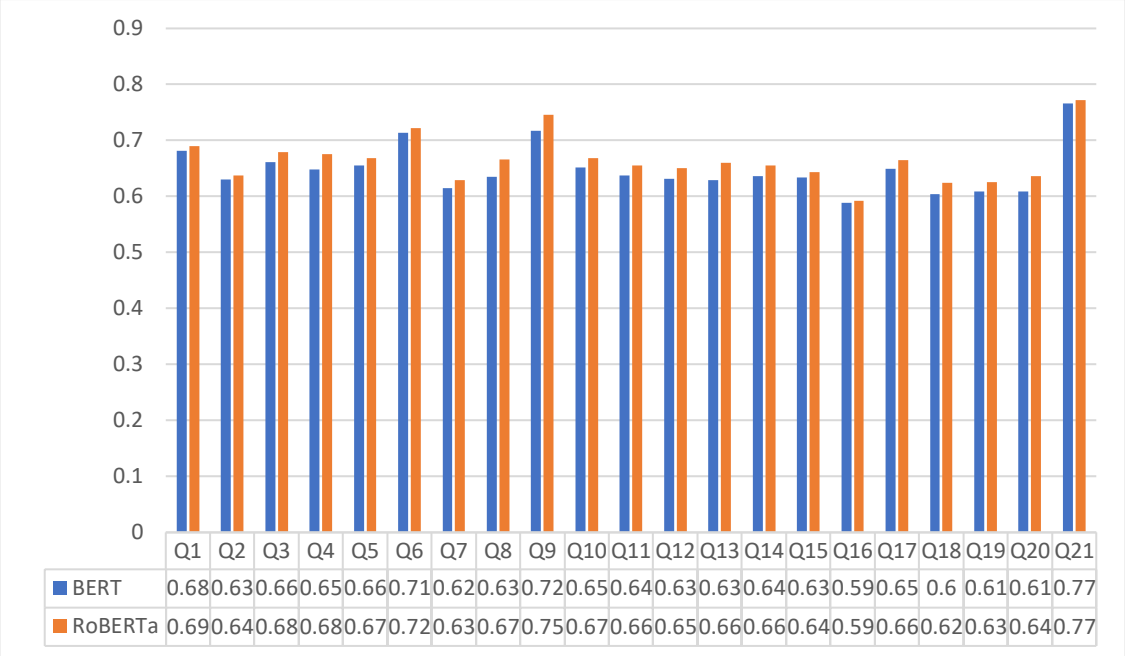


Figure 5: To decide the number of Epoch, we test our system on Q3-Q6

Later we use RoBERTa as the core of the system [18]. Since RoBERTa is optimized on the basis of BERT, and the authors have expanded the training dataset, trained with longer sequences, dynamically generated the shields used by MLM. We test them with our training data, and Fig. 6 shows that RoBERTa is significantly more accurate than BERT in our system.



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21
BERT	0.68	0.63	0.66	0.65	0.66	0.71	0.62	0.63	0.72	0.65	0.64	0.63	0.63	0.64	0.63	0.59	0.65	0.6	0.61	0.61	0.77
RoBERTa	0.69	0.64	0.68	0.68	0.67	0.72	0.63	0.67	0.75	0.67	0.66	0.65	0.66	0.66	0.64	0.59	0.66	0.62	0.63	0.64	0.77

Figure 6: Q1-Q21 accuracy of our system during development

3.3. Post-processing

In the first run, our system output the original prediction result as a baseline for the follow-up runs, there is no any weighting of the predicted results. The prediction of each question is a simple majority vote, the system output the answer with the largest cumulative number. Fig. 7 shows the prediction distribution of each question. We find that the prediction, affected by the training set, tend to favor less severity.

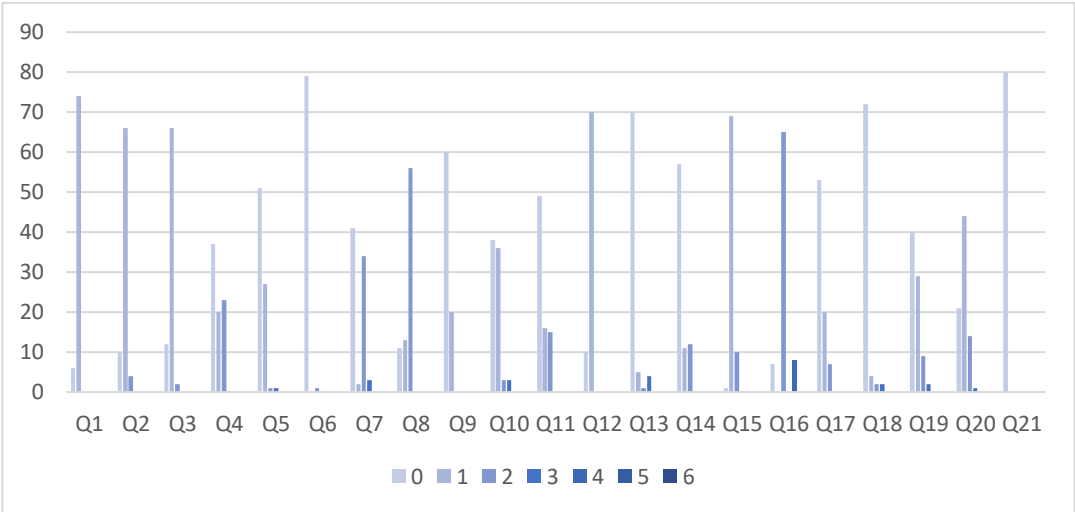


Figure 7: The answer prediction distribution of our Run1

In the second run, the predictions are adjusted according to the predictions in the first run, with a fixed-weight weighting mechanism to give with a higher severity answers. For example, Q1 has four different level of severity (from 0 to 3), so we give 1 to 4 as the severity weight. That is, if one posting

is predicted as 1 by our model, we count it twice; if one posting is predicted as 2 by our model, we count it 3 times; and if one posting is predicted as 3 by our model, we count it 4 times. Our system finally set the prediction of the question according to the maximum number after the weighting. Fig. 8 shows Run2 prediction distribution, the distribution is pushed to the higher levels of severity.

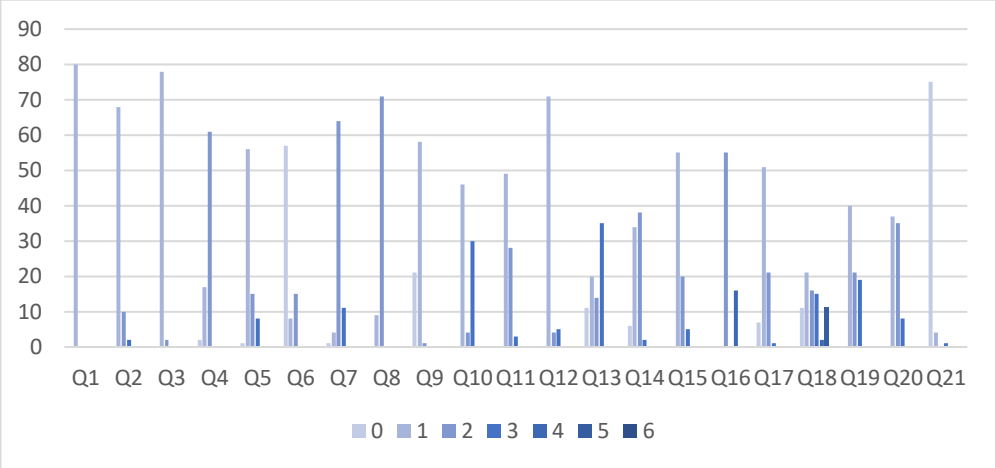


Figure 8: The answer prediction distribution of our Run2

In Run 3, the weighting is adjusted based on the percentage of the training data distribution. We use the percentage of distribution in Table 1 as the threshold value, the answer with the highest percentage of each question is selected as a default answer. Our system modify the weighting in order from the most severe to the slightest order, as long as the percentage of the prediction result is greater than the percentage of the distribution of training data, the prediction results as the final answer. For example, for Q1, the distribution in training data is (0:35%, 1:51%, 2:9%, 3:6%), the answer with the highest percentage is 1 in Q1, then 1 is our default answer. Suppose the original system predict output distribution for some user is (0:36%, 1:51%, 2:10%, 3:3%), our system will first check the percentage of answer 3, in this case 3% does not exceed the 6% threshold, so it is not our choice. Our system then will check the percentage of answer 2, in this case is 10%, which does exceed the 9% threshold, our system will output answer 2. In short, our system tends to choose the answer with higher severity.

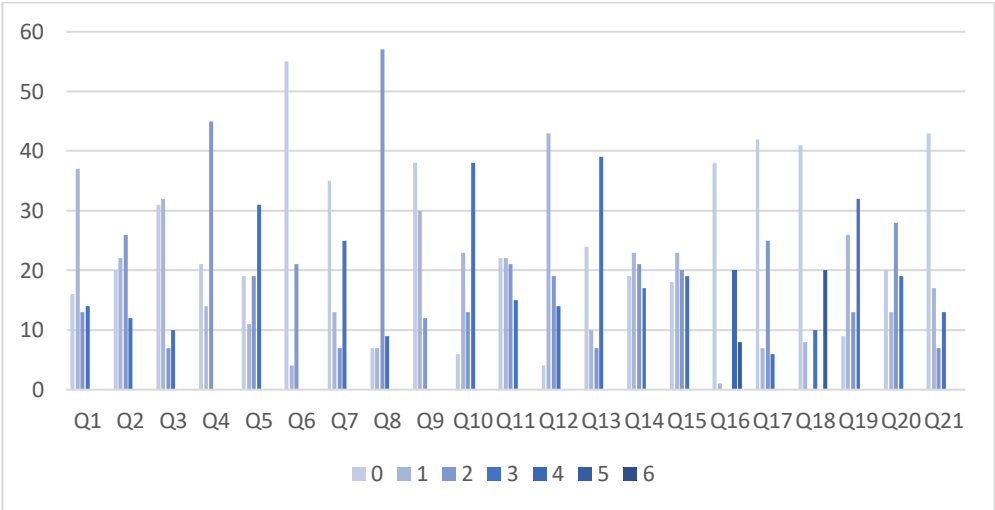


Figure 9: The answer prediction distribution of our Run3

Fig. 9 shows the Run 3 prediction distribution. Unlike the first two runs, this distribution results are more broadly, less concentrated. However, overall performance is not the best.

4. Results and Discussion

Table. 2 shows the results of our three Runs this year in task 3 [8], and we compare them to be best results in last two years [5] [9]. This year, nine teams sent out 36 runs, of which we got the best results in ADODL and DCHR, and our ADODL performed better than the best results in the past 2 years.

The best DCHR and AHR results in year 2019 still hard to match. The best DCHR's practices can be seen in [10], they use unsupervised methods to make the results. The authors also noted that by a simulation in Fig. 10 that comparing the results of random, the authors felt that although the data were the best, they did not perform better than random. The best AHR practice can be seen in [11], the authors first decided each user's depression level then decided the answer to the questionnaire. This approach is totally different from our approach.

Table 2
System performance of our runs and best results in recent three years

	AHR	ACR	ADODL	DCHR
CYUT RUN1	32.02%	66.33%	75.34%	20.00%
CYUT RUN2	32.62%	69.46%	**83.59%**	**41.25%**
CYUT RUN3	28.39%	63.51%	80.10%	38.75%
Best result in this year [8]	35.36%	**73.17%**	**83.59%**	41.25%
Best result in year 2020 [5]	38.30%	69.41%	83.15%	35.71%
Best result in year 2019 [9]	**41.43%**	71.27%	81.03%	**45.00%**

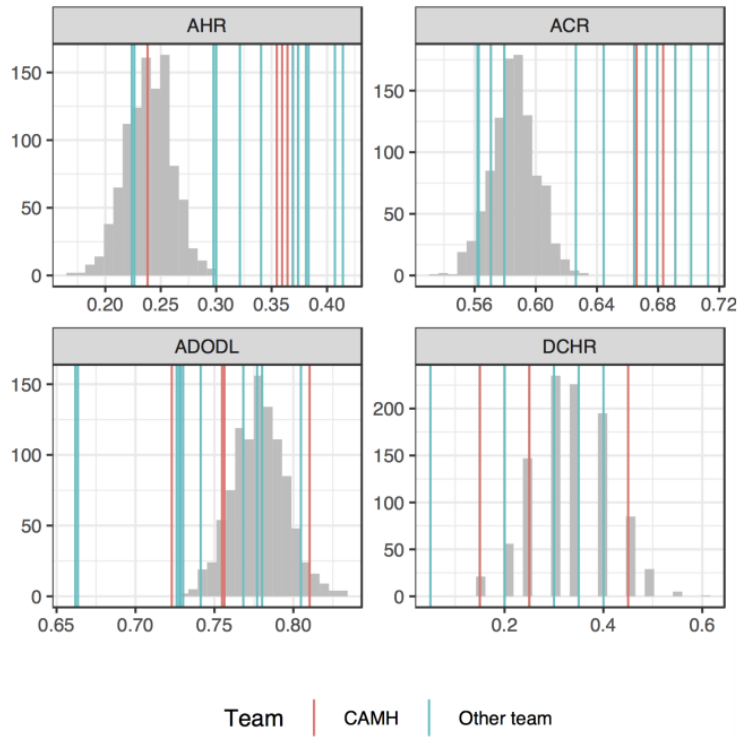


Figure 10: Histograms of randomly generated submissions with team submissions marked by vertical lines. (2019)[10]

Table. 2 shows that the overall performance of Run2 is better than the other two runs. We further show the number of correct prediction of the Q1-Q21 individual results in Fig. 10. From Fig. 11, we can see that Run2 give better prediction in 9 out of the 21 questions. We can also find that Q16 and Q18 are real hard to predict, where Q16's Run1 only predicts correctly once. This observation suggests that our weighting post-processing is valid, and most effective in Q16. However over-weighting also results

in a decrease in Run3 results, such that in Run3 of Q1, obviously it makes the prediction results worse. We find that the appropriate adjustment gives better results might due to the unbalanced distribution in training data. Adjustment to fit the training data distribution is an effective post-processing.

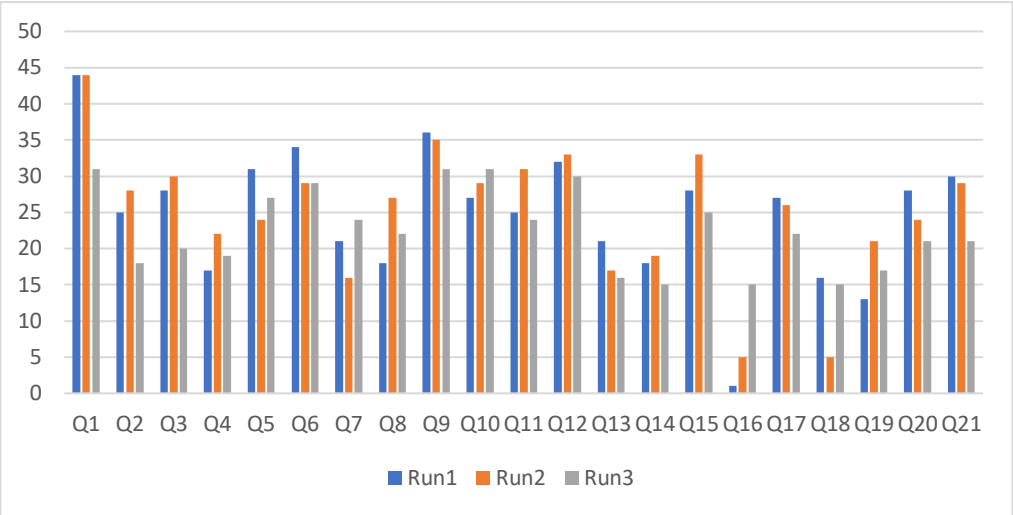


Figure 11: The number of correct prediction of the three runs

5. Conclusion and Future Works

The goal of eRisk T3 is to automatically assess the severity of depressive by analyze the postings of a person. We used a deep learning approach based on pre-trained model RoBERTa to build our system. We submit three runs with different post-processing weighting mechanism. Run2 gives the best ADODL and DCHR this year.

In our experiments, we assume that each posting will give the same information on the choice of the user. We believe that this is not a good assumption. Since a user might give positing in different emotions in different time, that will be very different from what the user might answer to each of the questions. This is one point that we will improve in the future. It should be that even for a user that shows higher level of severity according to the questionnaire, there will be only some of the sentences might show higher level of severity. Therefore, the sentences should be filtered with other tools. Only the ones that shows higher level of severity should be associated with the higher scores.

In the future, we plan to optimize the data, by comparing depression articles with non-depressive articles, extract the content of articles that shows depression, and remove the content of over-familiar articles, reducing the impact of useless content on the model.

6. Acknowledgement

This study was supported by the Ministry of Science and Technology under the grant number MOST 110-2221-E-324-011.

7. References

[1] SIMON KEMP.: DIGITAL 2021: GLOBAL OVERVIEW REPORT. URL:https://datar-eportal.com/reports/digital-2021-global-overview-report (2021).

[2] Eichstaedt, J.C., Smith, R.J., Merchant, R.M.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences (PNAS) 115(44), 11203–11208 (2018).

[3] Reece, A.G., Reagan, A.J., Lix, K.L.M. et al. Forecasting the onset and course of mental illness with Twitter data. Sci Rep 7, 13006 (2017). URL:https://doi.org/10.1038/s41598-017-12961-9

[4] CLEF eRisk: Early risk prediction on the Internet. URL:https://erisk.irlab.org/ (2021)
[5] Losada D.E., Crestani F., Parapar J.: Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview) (2020). URL:http://ceur-ws.org/Vol-2696/paper_253.pdf
[6] Chris McCormick.: BERT Word Embeddings Tutorial. URL:https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/ (2019)
[7] Use the huggingface pre-trained model to solve 80% of nlp problems. (2020, October 18). URL:https://www.bilibili.com/video/BV1Dz4y1d7am/?spm_id_from=333.788.videocard.15
[8] Parapar, J., Martin-Rodilla P., Losada, D. E., & Crestani, F. (2021, September). Overview of eRisk 2021: Early Risk Prediction on the Internet. In Proceedings of the Twelfth International Conference of the Cross-Language Evaluation Forum for European Languages (pp. tbp). Springer, Cham.
[9] Losada D.E., Crestani F., Parapar J.: Overview of eRisk at CLEF 2019 Early Risk Prediction on the Internet (extended overview) (2019). URL:http://ceur-ws.org/Vol-2380/paper_248.pdf
[10] Abed-Esfahani P., Howard D., Maslej M., Patel S., Mann V., Goegan S., and French L.: Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings. (2019). URL:http://ceur-ws.org/Vol-2380/paper_102.pdf
[11] Burdisso S.G., Errecalde M., Montes-y-G´omez M.: UNSL at eRisk 2019: a Unified Approach for Anorexia, Self-harm and Depression Detection in Social Media. (2019). URL: http://ceur-ws.org/Vol-2380/paper_103.pdf
[12] MADANI A., BOUMAHDI F., BOUKENAOUI A., KRITLI M.C., and HENTABLI H.: USDB at eRisk 2020: Deep learning models to measure the Severity of the Signs of Depression using Reddit Posts. URL:http://ceur-ws.org/Vol-2696/paper_39.pdf
[13] Martínez-Castaño R., Htait A., Azzopardi L., Moshfeghi Y.: Early Risk Detection of Self-Harm and Depression Severity using BERT-based Transformers iLab at CLEF eRisk 2020. (2020). URL:http://ceur-ws.org/Vol-2696/paper_50.pdf
[14] Maupomé D., Armstrong M.D., Belbahar R.: Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. (2020). URL:http://ceur-ws.org/Vol-2696/paper_53.pdf
[15] Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. p. 2968–2978. Association for Computational Linguistics (2017).
[16] Trifan A., Salgado P., Oliveira J.L.: BioInfo@UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. (2020). URL http://ceur-ws.org/Vol-2696/paper_43.pdf
[17] Jacob Devlin; Ming-Wei Chang; Kenton Lee; Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, (2019).
[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Computing Research Repository, (2019). arXiv:1907.11692. version 1