CYUT at eRisk 2022: Early Detection of Depression Based-on Concatenating Representation of Multiple Hidden Layers of RoBERTa Model

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

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

Abstract

Depression has been seen as a global crisis, with hundreds of millions of people around the world suffering from it. By analyzing people's writings on social media, a system has the opportunity to detect depression and can alert the person to seek medical help. Our team participated the CELF 2022 eRisk Task 2: Early Detection of Depression, a mission designed to detect people early for depression tendencies. Our research methodology focuses on improving the pre-training model RoBERTa. We ran a total of five experiments this year. The first one is regarded as a baseline using the pre-trained language model. Experiment two is to extract the output of hidden layers as a new representation. Experiment three is to obtain keyword features by extracting two categories of single word features. Experiment four is to train two models for the title and text separately, and integrate the results to make predictions. Experiment five is to integrate the methods of experiment two and experiment four. According to the results of the task evaluation, the method of experiment two is indeed better than using the pre-trained model. Experiments 4 and 5 performed well on the Task's Ranking-based evaluation after testing 1000 writings.

Keywords

Deep Learning, RoBERTa, Depression Detection

1. Introduction

Depression is a popular disease of civilization in the 21st century, and according to data published on the website of the World Health Organization (WHO), 264 million people worldwide is suffered from depression in 2020. Nowadays, people are accustomed to sharing their living through social media, and analyzing these posts can observe the depression tendencies of the authors. In 2018, Eichstaedt, J.C., Smith, R.J., Merchant, R.M et al. used messages in Facebook posts to predict depression in their medical records [1]. In 2017, Reece, A.G., Reagan, A.J., Lix, K.L.M et al. used Twitter data to predict the onset and course of mental illness [2]. The eRisk organizers have organized related tasks in the CLEF lab. Last year's CLEF eRisk Task 3: Measuring the severity of the sign of depression [3] was also used posted writings on social media to predict the user's severity of depression. We have also participated in this lab in 2021 [10], using the social media corpus to train BERT [11, 12] and RoBERTa [7], and weighted the predictions of each post to calculate the degree of depression of the authors. From the overview of eRisk at CLEF 2021 results report [13], we learned that our approach performed well, and we found that each team has its own methodology. Alhuzali et al. team ran three different pre-trained language models to observe the practicality and strengths of each model, and it was learned from their experiments that the pre-trained model was trained for sentiment analysis, which helped to strengthen the model's judgment of the degree of depression [14]. Inkpen et al. team proposed two main approaches, the first approach is using of pre-trained models, and by analyzing the relevance of all posts

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
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)

and BDI answers, they believed it should be noted that not all categories were discussed in posts [15]. The second approach is to classify posts in different topics, and find the most relevant topics through the word vectors with the corpus. Bucur et al. team and Spartalis et al. team also used the pre-trained model approach [16,17], the difference being that one was trained to analyze post similarities and the other was to analyze feature-based transfer learning.

To analyze people's psychological conditions through a wide range of information in social media is widely appreciated. CLEF eRisk also gave three different tasks this year [4], namely Task 1: Early Detection of Signs of Pathological Gambling, Task 2: Early Detection of Depression, Task 3: Measuring the severity of the signs of Eating Disorders. Our team is involved in Task 2, a task designed to detect people for depression tendencies. The eRisk server iteratively provides user writing to the participating teams by releasing data step by step. How to diagnose the tendency to depression early through some data is part of the evaluation indicator, that is, the evaluation not only considers the correctness of the system output, but also considers the time point at which its decision is published.

2. Data and Pre-processing

The data used in this paper is the dataset provided at eRisk 2022 Task 2: Early Detection of Depression [5][18]. The data contains text from multiple users, each of whom typically provides a large amount of written text in the XML format as in Figure 1. ID: Contains the anonymous ID of the user, title: the title of the post (keep blank for comments), INFO: the source of the post, TEXT: the content of the post or comment.

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

Figure 1: The XML format of each post in the dataset [5], where ID is the anonymous user ID, TITLE is the post title, INFO is the source, and TEXT is the content of the post

The Early Detection of Depression datasets are listed in Figure 2. There are datasets in 2018 and 2017 respectively, each is collected social media posts of that year, and are divided into two categories: depression (pos) and non-depression (neg). This paper uses 2018 data set for model training, and the 2017 data set for verification.

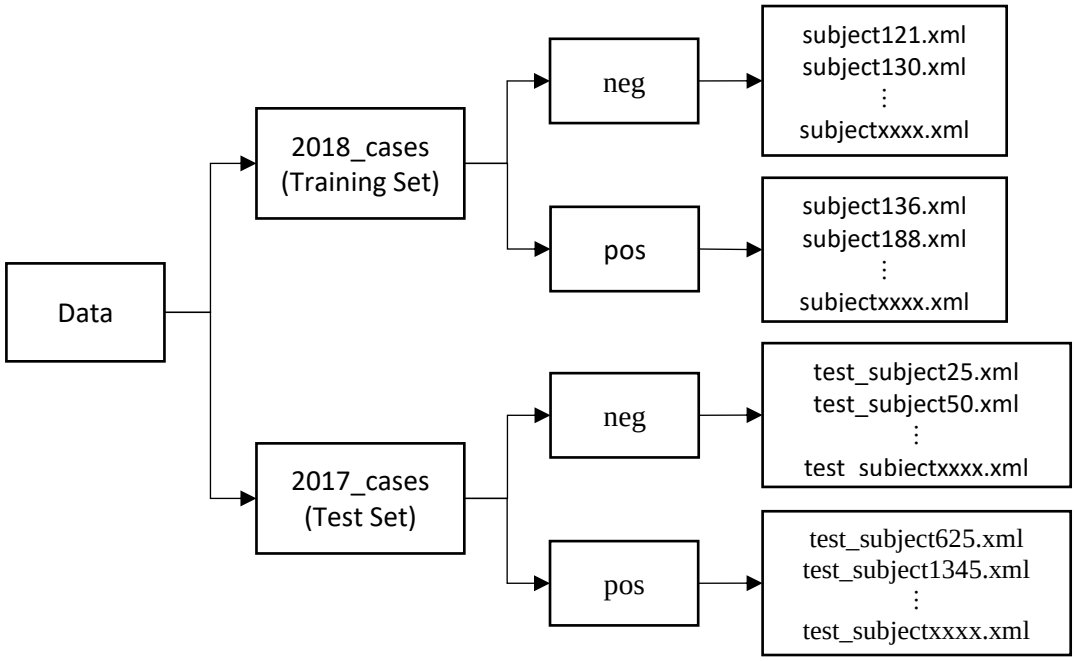

Figure 2: Data sets provided by the eRisk organizers [5]. We use the data in year 2018 as the training set and the data in year 2017 as the validation set during system developing.

Since the dataset was collected from the forum and has not been processed, it contains paths, URLs, some special characters, and so on. Therefore, we use regular expression do the preprocessing on the title and text of each document as shown in Figure 3. Special characters, paths, URLs, parentheses, and punctuation are removed. The number of training and verification after preprocessing is as shown in Table 1.

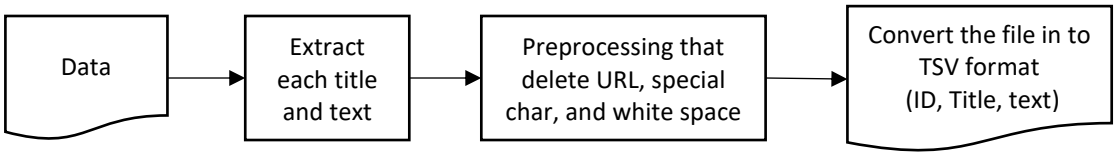

Figure 3: Data set normalization

Table 1
Number of training and validation

	Pos(People)	Neg(People)	Pos(posts)	Neg(posts)
Training Set	79	741	40,385	498,004
Test set	52	349	17,431	157,433

The training materials came from a total of 820 people, of which the majority (741 people) were non-depression, which shows that the data is extremely unbalanced. The situation is often encountered in real world problems, how to effectively filter the post is an important issue, and it is also the main consideration of our research. According to the previous observation [6], there is a difference between the length and amount of words used. We show them in Figure 4 and Figure 5, respectively, for the length of the text and the number of words. Blue represents the post of non-depression ones, red represents post written by depression users, and the X axis is the total number of posts 538,389. The Y-axis indicates the length of the post and the number of words, respectively, and statistically there are indeed some posts that show that the posts by non-depression users have a longer length and more words than the posts by depression users. Therefore, according to this data, we removed the posts with

length more than 1000 and the number of words over 500. But this distinction is still limited, and most of the posts are still similar in length to the number of words.

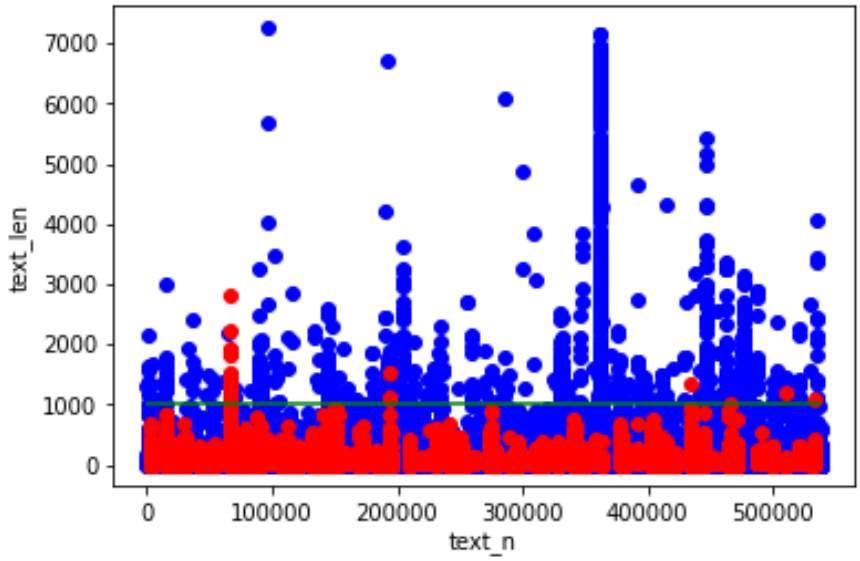

Figure 4: Statistics on the length of each post (text_n: number of posts, text_len: length of post).

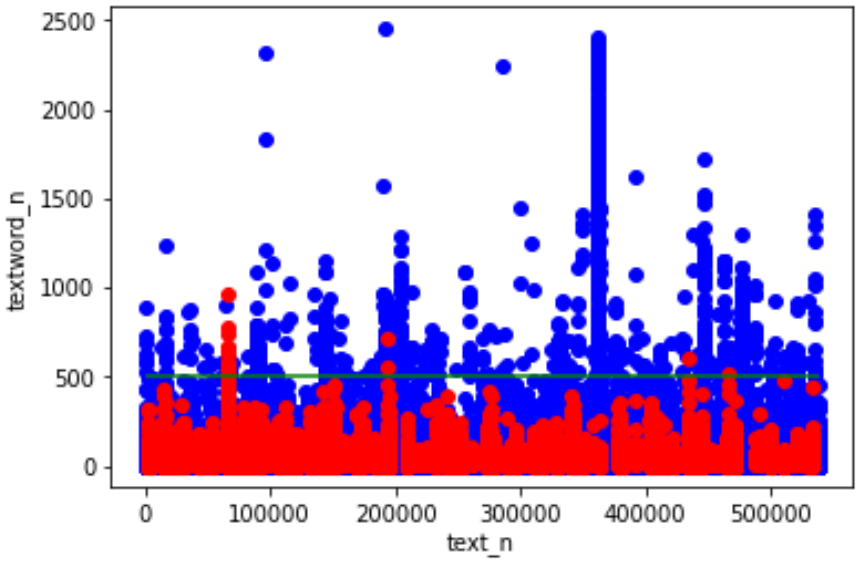

Figure 5: Statistics of the number of words in each post (text_n: number of posts, text_len: number of words)

3. Our Approach

We describe our system settings in sub-section 3.1 and how we evaluate our system in sub-section 3.2. The experiment settings of our 5 runs is shown in the following 5 sections.

3.1. Operating environment and model parameter settings

Model is trained on Google Colab Pro, the training data is listed in Table 1, the data is divided into 80% for training, 20% verification, tokenizer and model are roberta-base. The hyper parameters settings

are: max length is set to 128, batch size is set to 100, hidden size is set to 768, learning rate is set to 1e-5, weight decay is set to 1e-2, and epoch of fine-tuning is set to 2.

3.2. Evaluation model method

Evaluation processes is shown in Figure 6, there are two evaluation modules. One is to predict the outcome of each data's depression tendency, and the other is to predict whether the statistical prediction results are for the corresponding person to determine whether there is a depression tendency. The process is to give test set data to the experimental model to predict whether it is a tendency to be depressed, and calculate the model Precision, Recall, and F1-score scores. And the data results are statistically judged by the corresponding person, adjusted from 1% to 99% of the symptomatic data, and calculate the Precision, Recall and F1-score scores under different proportions to find the best F1-score score.

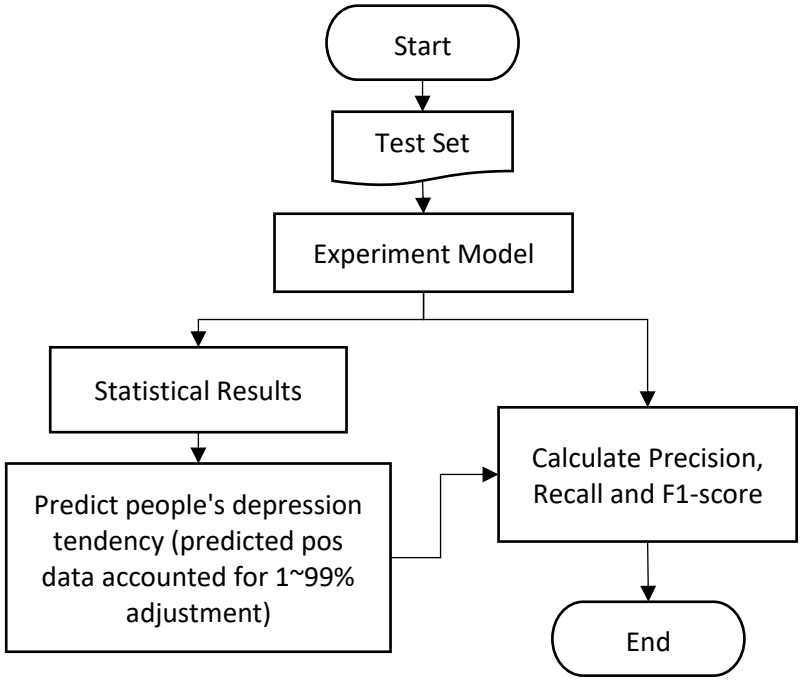

Figure 6: Evaluation process

3.3. Experiment 1: RoBERTa

The pre-trained RoBERTa model [7] was used as a baseline model for evaluating model score changes against subsequent comparisons. The flowchart of experiment one is shown in Figure 7, the only preprocessing is focus on the data imbalance issue. The treatment is to reduce the number of posts extracted from the documents of the non-depression people in the training set (up to 500 posts per person). The total training posts are 268,866 and use the TEXT part for model training.

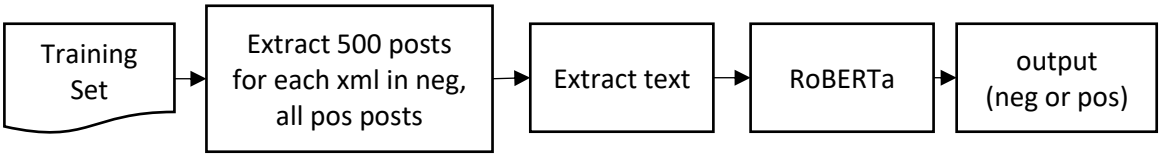

Figure 7: Experiment 1 process

3.4. Experiment 2: RoBERTa (Extract output of hidden layers)

The main idea of experiment 2 is to change the embedding representation of an input sentence in the RoBERTa model. The first tokens of each of the last four hidden layers are extracted from the model for improvement [8]. This token represents the corresponding output vector of each layer, which means that this token is the result of the model's representations in each hidden layer. In this experiment, the results given by the last layer vector will be used for linear classification (see Figure 8). We want to know if the model prediction can be improved by extracting multiple output vectors.

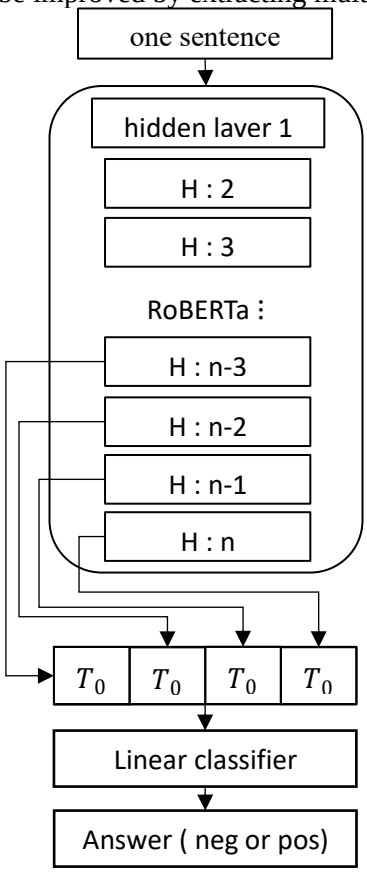

Figure 8: Extracts the output vector from the last four layers of the model's hidden layer and joins the four output vectors as the input vector of the linear classifier.

3.5. Experiment 3: Building feature dictionary + RoBERTa

The experiment 3 setting is not to unconditionally discard the non-depression data, but to filter the data by creating a feature dictionary (to retain the information that can be matched by the dictionary). From the progressive reference [9], when people's thoughts and emotional reactions are different the usage of words will be different too, that is why the negative emotion dictionary has been used in the past. However, since that it is easy to publish posts using social media, new buzzwords or lists may be generated at any time. Therefore; we try to extract a new dictionary of features by comparing the posts from depression users and non-depression users.

3.5.1. Build feature dictionary

The extraction process is shown in Figure 9, the frequency of words in training data from depression and non-depression users are counted separately. Some words only appear a few times, such as: personal names, place names, song names, etc., so two threshold values (5, 16) are set for the frequency of occurrence of words with depression and non-depression users. Two feature dictionaries are extracted, and the number of words in the characteristics of the non-depression is 19,214, and the number of words with the depression is 1,106.

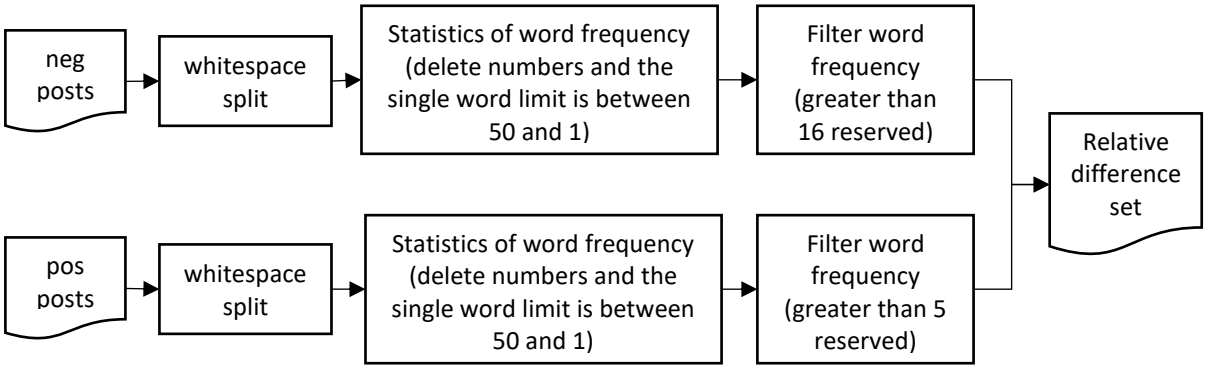

Figure 9: Building feature dictionary

3.5.2. Experiment process

The flow chart of experiment 3 is shown in Figure 10, training data are screened by matching with a feature dictionary. After screening, a total of 129,544 non-depression data were screened, and the depression data was also screened for the purpose of strengthening the training of these data, a total of 902 cases. The processed training data contained all posts from depression users (40,353 posts) and more characteristic posts (129,544+902 posts) after screening, with a total of 170,799 posts. The training data is used to fine-tune the RoBERTa model.

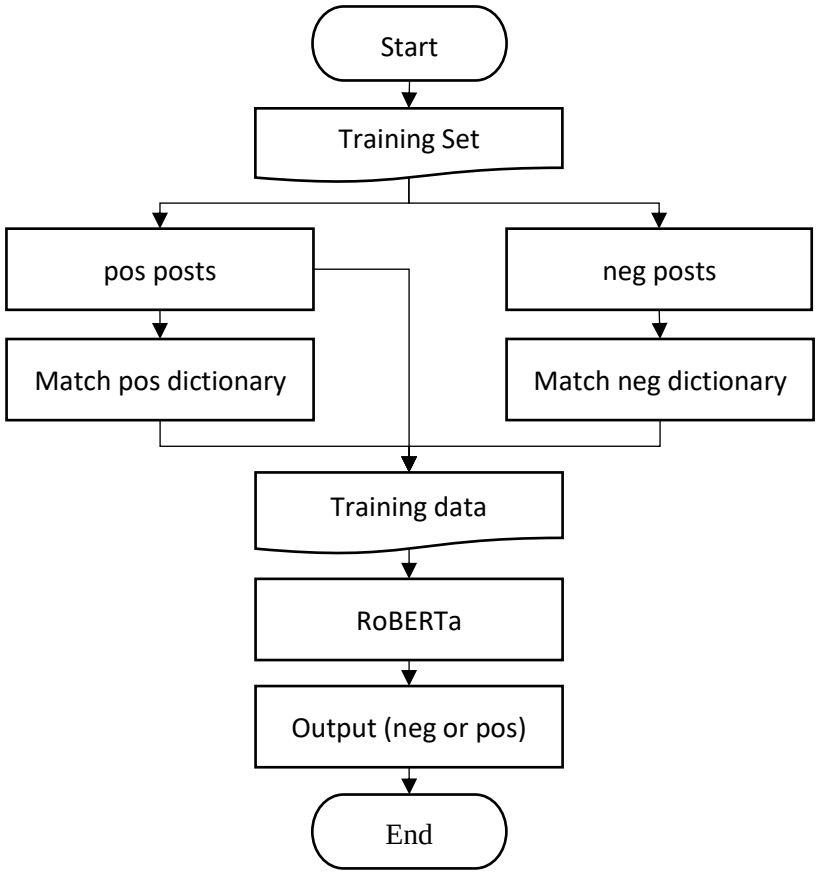

Figure 10: Experiment 3 process

3.6. Experiment 4: Combining Title and Text Prediction Models

Experiment 4 is to train two models for the title and for the text separately. According to the observation of the dataset, some of the data is only containing the title and no text, and this experiment design is to deal this situation. Experimental process is shown in Figure 11, the system extracts the title and body of each post from the training data, and the title and body each haw a separate RoBERTa model for training, and the results are integrated to make judgments by a linear classifier.

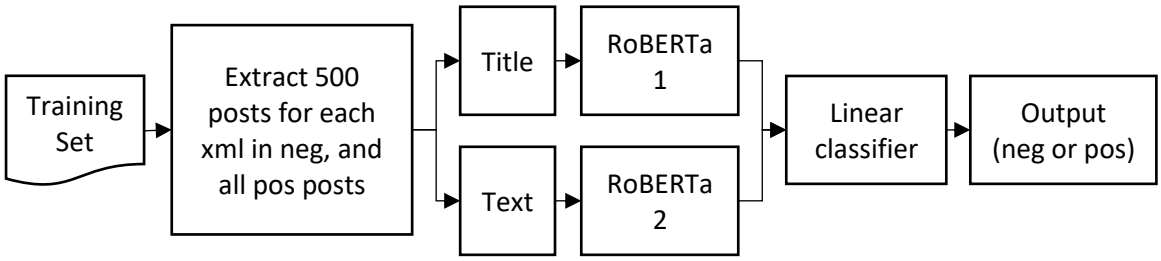

Figure 11: Experiment 4 process

3.7. Experiment 5: Combining experiments 2 and 4

We observed from the validation evaluation results of experiment 2 (Table 2) that the method of extracting information from the hidden layer is effective, so we improved the process of experiment 4 according to the method of experiment 2 for experiment 5.

4. Results and Discussion on System Development and eRisk task 2

The result of our five experimental processes according to the Section 3.2 evaluation methodology. Verify using the 2017 dataset test data. The first assessment is to determine whether there is a depression tendency result is shown in Table 2, which is the result of the 401 users. According to the data results, we find that the decision proportion of depression posts is predicted to be different on whether the user has a tendency to be depressed. Figure 12, Table 3 are the metrics for all experiments at the proportion of the best F1-score score.

Table 2
The evaluation experiment judges the results of each data

Run	Precision	Recall	F1-score
Experiment 1	30.94%	12.62%	17.93%
Experiment 2	32.55%	13.38%	18.97%
Experiment 3	10.54%	60.92%	17.98%
Experiment 4	30.21%	12.27%	17.46%
Experiment 5	26.77%	9.77%	14.32%

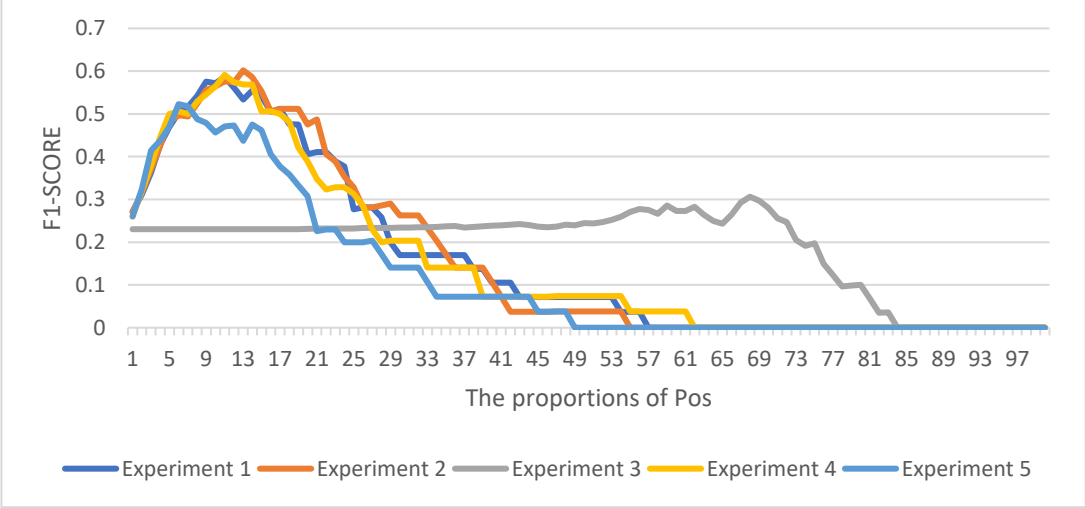

Figure 12: Changes in F1-score under different proportions of Pos. Where y-axis is the F1-score, and the x-axis is the proportions of posts that the system predicts as Pos. This figure shows that our model performs well when it finds that about 15% of a user's post is depression.

Table 3
The best results of an evaluation experiment for predicting people's tendency to depression

Run	Pos Percentage	Precision	Recall	F1-score
Experiment 1	11%	53.12%	65.38%	58.62%
Experiment 2	13%	60.78%	59.61%	60.19%
Experiment 3	68%	28.81%	32.69%	30.63%
Experiment 4	11%	53.96%	65.38%	59.13%
Experiment 5	6%	39.60%	76.92%	52.28%

4.1. Experiment 2: Discussion

The evaluation results show that extracting multiple output vectors can effectively improve the performance, which is more accurate than using only the last layer to predict the results. The result of experiment 2 in Table 2 is more outstanding than experiment 1 in most evaluation matrices. From Figure 12, it can be seen that the best result of F1-Score is 60.19% when the proportion of depression in this experimental model is 13%. In the comparison of Table 3 evaluation results, the F1-score of Experiment 2 is 2% better than that of Experiment 1.

4.2. Experiment 3: Discussion

The results of the assessment did not succeed in improving the prediction. The sharp increase in Recall was accompanied by a sharp decline in precision, which made it easy to make mistakes in judging the tendency to predict depression. As shown in Table 3, accuracy, recall, and F1-score were among the worst of all experimental evaluations. The main reason for this situation is that the data is over-screened. In the establishment of feature dictionaries, too extreme methods are taken, and only words that appear in one of the categories are retained, which also leads to excessive exclusion of training materials. This in turn leads to insufficient model training. From Figure 12, it can be observed that the training model effect is very poor, when the proportion of symptoms is greater than 68%, the F1-score is greatly reduced. This is abnormal, it means that the model tends to predict that there is depression tendency result, but in fact, the number of people with non-depression tendency is much greater than the number of people with depression tendencies. This condition, as mentioned earlier, is due to over-excluding the results of the training data. And because the number of depression data is too rare after

matching, and all the data on depression tendency are put back to the training data, this also leads to the training of the model to have a bias toward predicting depression.

4.3. Experiment 4: Discussion

The evaluation results have not improved significantly, and it can be observed from Table 2 that the evaluation results of experiment 4 and experiment 1 are not much different, and only about 0.5% are improved in judging whether people have a depression tendency. This is slightly helpful, but the effect is not as effective as experiment 2. However, compared with the previous experiments, this model can predict the results of the title without the text, so it has different applications similar to the previous experimental models.

4.4. Experiment 5: Discussion

The results showed that instead of improving, they deteriorated, such as the results of experiment 4 in Figure 12, which were better than the evaluation results of experiment 5. And the reason for this might be too much different information, and finally the model is difficult to converge and make wrong judgments. From experiment 5 we found that the effect of this approach is limited, the vector size of RoBERTa hidden layer is 768, so the last four hidden layers a total of 6144 dimension vectors will be used. However, it might cause difficult to converge the results for a linear classification, so the model judgment ability is reduced.

Table 4
Decision-based evaluation

Run	P	R	F1	$ERDE_5$	$ERDE_{50}$
Experiment 1	0.165	0.918	0.280	0.053	0.032
Experiment 2	0.162	0.898	0.274	0.053	0.032
Experiment 3	0.106	0.867	0.189	0.056	0.047
Experiment 4	0.149	0.878	0.255	0.075	0.040
Experiment 5	0.142	0.918	0.245	0.082	0.041

Table 5
Presents the ranking-based results[18]

	1 writing			100 writing			500 writing			1000 writing		
Experiment	P@10	$NDCG@10$	$NDCG@100$	P@10	$NDCG@10$	$NDCG@100$	P@10	$NDCG@10$	$NDCG@100$	P@10	$NDCG@10$	$NDCG@100$
1	0.50	0.49	0.37	0.50	0.52	0.54	0.60	0.59	0.58	0.70	0.72	0.61
2	0.70	0.77	0.37	0.60	0.72	0.58	0.60	0.72	0.61	0.70	0.80	0.62
3	0.00	0.00	0.16	0.10	0.07	0.25	0.10	0.19	0.31	0.10	0.12	0.29
4	0.10	0.07	0.12	0.70	0.70	0.57	0.70	0.72	0.59	**0.80**	0.74	0.60
5	0.10	0.06	0.12	0.60	0.68	0.55	0.60	0.69	0.59	**0.80**	0.84	0.61

4.5. Formal Results in eRisk 2022 Task 2

We ran the above five experimental models on this Task 2, processing a total of 2000 iterations of user writing, which took 7 days and 12 hours to complete. The Decision-based evaluation results were not particularly pronounced (Table 4), and the Recall was on the high side of each experimental model.

We believe that the reasons for this result are due to the different way of evaluation. During the system development phase, all of the writing are given at once. While the task is to give each user a writing post at a time in an iterative way, and predict the data the user's depression tendencies early. However, we have a good performance in the ranking-based results, and from Table 5, we can observe that the more information our model gets, the evaluation score continues to rise, and P@10 the best performance out of 1000.

5. Conclusions

During the system developing phase, we use all of the user's writing training to train the model, which is different from task 2, which is to give one post at a time in an iterative manner. Therefore, our model is weaker in early detection of user depression, but has a good performance in the ranking-based results. Compared with the baseline of our experiment one, the results of experiment three are not as expected. We learned from it that the statistical common word count ratio as a classification feature might be overfitting. Extracting the output vector in the hidden layers in Experiment 2 as a new representation has indeed been effectively improved, and it is more accurate to predict the result than experiment 1 directly using a pre-trained model. In the design of experiment four, we combined the model trained on the body text with the model that trained on the titles. Although this method has not significantly improved in the evaluation effect, but compared to only the body text, the combined model can handle special cases that title is missing or body text is missing.

6. References

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

[2] Reece, A.G., Reagan, A.J., Lix, K.L.M. et al.: Forecasting the onset and course of mental illne ss with Twitter data. Sci Rep 7, 13006 (2017). URL:https://doi.org/10.1038/s41598-017-1296 1-9.

[3] CLEF eRisk: Early risk prediction on the Internet, 2021. URL: https://erisk.irlab.org/2021/ind ex.html

[4] CLEF 2022 Conference and Labs of the Evaluation Forum, 2022. URL: https://clef2022.clef-i nitiative.eu/index.php?page=Pages/labs.html#erisk

[5] eRisk 2022 Text Research Collection, 2022. URL: https://erisk.irlab.org/eRisk2022.html

[6] Fidel Cacheda, Diego Fernández, Francisco J. Novoa, Víctor Carneiro.: Analysis and Experim ents on Early Detection of Depression, 2018. URL: http://ceur-ws.org/Vol-2125/paper_69.pdf

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mi ke Lewis, Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Computing Research Repository, (2019). arXiv:1907.11692. version 1

[8] Chris McCormick, Nick Ryan.: BERT Word Embeddings Tutorial, 2019. URL: https://mccor mickml.com/2019/05/14/BERT-word-embeddings-tutorial/

[9] Yen-Shuan Huang, Wen-Hsiang Lu.: Predicting Web User's Tendency of Depression Using Negative Thought-Driven Depression Model, 2015. URL: https://hdl.handle.net/11296/uskn27

[10] Shih-Hung Wu, Zhao-Jun Qiu.: A RoBERTa-based model on measuring the severity of the signs of depression, 2021. URL: http://ceur-ws.org/Vol-2936/paper-86.pdf

[11] 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).

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin.: Attention IsAll You Need. arXiv:1706.03762v5 6 Dec 2017.

[13] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, Fabio Crestani.: Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview), 2021. URL: http://ceur-ws.org/Vol-2936/paper-72.pdf
[14] Hassan Alhuzali, Tianlin Zhang, Sophia Ananiadou.: Predicting Sign of Depression via Using Frozen Pre-trained Models and Random Forest Classifier, 2021. URL: http://ceur-ws.org/Vol-2936/paper-73.pdf
[15] Diana Inkpen, Ruba Skaik, Prasadith Buddhitha, Dimo Angelov, Maxwell Thomas Fredenburgh.: uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning, 2021. URL: http://ceur-ws.org/Vol-2936/paper-79.pdf
[16] Ana-Maria Bucur, Adrian Cosma, Liviu P. Dinu.: Early Risk Detection of Pathological Gambling, Self-Harm and Depression Using BERT, 2021. URL: http://ceur-ws.org/Vol-2936/paper-77.pdf
[17] Christoforos Spartalis, George Drosatos, Avi Arampatzis.: Transfer Learning for Automated Responses to the BDI Questionnaire, 2022. URL: http://ceur-ws.org/Vol-2936/paper-84.pdf
[18] Javier Parapar, Patricia Martín-Rodilla, David E. Losada, Fabio Crestani.: Evaluation Report of eRisk 2022: Early Risk Prediction on the internet, In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022. Springer International Publishing, Bologna, Italy. 2022.