HULAT-UC3M at Task1@eRisk 2025: Detecting Depression Using Machine Learning Approaches

Javier Campos-Molina[1], Paloma Martínez[1]

[1]Computer Science and Engineering Department, Universidad Carlos III de Madrid, Leganés, Madrid, Spain

Abstract

This paper describes the participation of HULAT-UC3M research group at Task 1: Search for Symptoms of Depression at eRisk 2025 shared task [1]. A proposal composed of three steps is proposes. The first is to train a SVM multi classifier using the embeddings from all-MiniLM-L6-v2 pretrained model to classify all the sentences into their corresponding symptom. Second step consists on a filter to select the most representative 1000 sentences to be sent and finally we will get the score for the sentences chosen in the previous step using a rule-based model and a encoder-based transformer (RoBERTa) for sentiment analysis. Performance of the best model is NDCG of 0.053 and P@10 of 0.157.

Keywords

Natural Language Processing, Machine learning, Depression detection, Classification, LLM

1. Introduction

Depression is a problem in today's society. According to the World Health Organization (WHO)[2], 3.8% of the world's population suffers from depression, 5% in adults and 5.6% in people over 60 years of age. It also affects approximately 50% more women than men. However, the most severe problem is not the depression itself, but what triggers it. Every year around 700,000 people commit suicide [3] and it is the fourth leading cause of death among 15-29 year old. In many countries no attention is paid to this type of illness and even more than 75% of people in low and middle income countries do not receive care.

Early detection and treatment in users with symptoms of depression is essential to improve the quality of life of people and avoid suicides. The participation of the team in eRisk task [1, 4] is in order to learn new mechanisms and possible solutions to this big problem in actual society using modern approaches with machine learning.

2. Related Work

Starting from 2019 we have one of the participants using a text classifier called SS3 [5] [6] for solving task 3 [7]. The task is related, but it is not exactly the same like the one solved in this paper as it consists of classification of depression severity instead of classifying in symptoms and scoring them. The classifier previously mentioned, SS3, is a probabilistic model using statistic in order to associate some words. For each word creates a probability of being associated with other words taking into account if it appears previously together with that word or not. Its important to mention that this is not a transformer although it may appear similar in the sense that it assigns a probability between 0 and 1 to each word in relation to others. SS3 does not use self-attention as transformers does, but it relies on probabilistic functions such as confidence (cf), support (sf), and credibility (cv) to model context.

In 2020 edition [8] some systems proposed solutions based on roBERTa model, a model that is more powerful than BERT as they were trained with times 10 more data. The model has a tokenizer itself that was used to create the tokens, then create the embeddings and finally they did the classification

CLEF 2025: Conference and Labs of the Evaluation Forum, September 09–12, 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

100472233@alumnos.uc3m.es (J. Campos-Molina); pmf@inf.uc3m.es (P. Martínez)

0000-0003-3013-3771 (P. Martínez)

and a softmax as last layer to compute the probabilities. The team using this approach was the best in terms of accuracy with more than a 69% [8].

In 2021 edition [9] some of the proposed systems followed similar approaches to those of 2020. One of the studies used BERT and roBERTa together [10] and the comparison of results were in favor of roBERTa, as it was expected for the reason previously mentioned. Other systems proposed different probabilistic methods similar as SS3 back into 2019. In this case, one of the groups participating on the task proposed a system using Latent Dirichlet Allocation (LDA) [11], that consists of a Bayesian network, in combination to sentence transformers and classical classifiers. LDA is very popular in unsupervised learning tasks. In other hand and although it is not a task related to depression but to self-harm we have some projects using interesting systems [9]. Self-harm task was a binary classification task between users that needs to classify them by the ones at some point have harmed themselves and the ones who have not. One of the team [12] used Yake for one of their runs, that is a model that takes the most important words out of a sentence but it did not work as expected as it removed the important signs of self-harm from the sentence. An additional run using VADER [12] model but there are no results available for this model as the team did not submit the run.

In 2022 edition [13], the system described in [14] used roBERTa but in addition they used a model called MiniLM that is a model derived from roBERTa and BERT architectures, with whole self-attention but it is a multilingual model and is able to work in different languages. RoBERTa model is only specialized in English texts. Also MiniLM is smaller and faster than roBERTa and could help in terms of efficiency. Another team introduced a fully connected neural network (FCNN) in the third run combined with previously used systems as support vector machine (SVM) and transformers. The results were very good in terms of recall (0.816 compared to 0.745 of the first team) but not in terms of precision (0.283 compared to 0.682). The team that won the competition [15] used the bag of words (BOW) approach, combined with entropy-based weighting and a SVM classifier. BOW is a technique for converting text into numerical representations, enabling the use of classical machine learning models as SVM, used in this system. This team applied TF-IDF weighting enhanced with entropy, giving higher importance to the most relevant words while reducing the influence of frequently occurring but not interesting terms. Additionally, they employed chi-square feature selection to further improve classification speed performance by retaining only the most relevant terms from the previous step by reducing the total amount of data.

eRisk 2023 edition [16] changed from binary to multiclassification classification. The objective of the task is to classify symptoms of depression according to BDI-II Questionnaire [17] and give them an score from 0 to 10. With the rise of generative AI, some of the teams [18] starting using LLM in order to generate more data as is it done in other approaches outside this task [19][20]. One of the teams [18] used ChatGPT in order to generate more data using a prompt for each of the symptoms. Then it was combined with some new models that performed well capturing semantic relations as MentalRoBERTa but the results were not good in comparison to the first team in the competition. The max average precision out of the 5 runs were 0.104 far below 0.319. Another team [21] also attempted to compare sentences based on their similarity by computing sentence embeddings using transformer-based models. However, due to the high computational cost of encoding all sentences, they first used the BM25 model [21] as a lightweight filtering step. Only the top ranked sentences, the most similar to the ones from the BDI-II questionnaire, were retained and then processed with the transformer models for similarity evaluation. The results were not good as the maximum AP of their runs is 0.039. Furthermore, the winners of 2023 year's task [22] used word2vec embeddings to capture semantic and grammatical similarities between words. Then, a soft cosine similarity is applied to compare each sentence in the dataset with the individual sentences describing each of the 21 symptoms from the BDI-II questionnaire. Specifically, if a symptom in the BDI-II is described by four different sentence options, the similarity between a dataset sentence and each of these options is computed individually, resulting in four similarity scores. These scores are then weighted using predefined weights for each option. The weighted similarity for each option is obtained by multiplying the similarity score by its corresponding weight. Finally, the total similarity between the sentence and the symptom is calculated as the sum of all the weighted similarities.

3. Method and system description

Before the development of the solution, an analysis of the labeled data given by the eRisk organization [4, 1] was done to check that there is enough data to compute a model and checking that all the labels were in the correct format. The main objective of the task is to get the score for each of the symptoms for 1000 sentences given in a test dataset. The process is structured in three steps. The first one is to train a multi classifier to classify all the sentences into their corresponding symptom. After that, the sentences were filtered according to different criteria for the different runs and finally, we will get the score for the sentences chosen in the previous step using different methods.

3.1. Multi classification of the sentences

For the multi classification problem, the model is trained using the annotated data from 2024 [4, 1], which are the ones offered for training. Unanimity, which means that all annotators agree to label that sentence as relevant to a particular symptom, is provided by organizers. Additionally, majority is also provided. The computed model uses the unanimity sentences because using the majority dataset could introduce sentences that increase the noise as not all of the annotators were agree. This can lead to misclassifications in the model for some of the sentences, even though they have more training data.

The proposed system uses a classical machine learning model, a support vector machine (SVM) [1] to do the multi-classification on the 21 symptoms present in the BDI-II questionnaire [17]. Taking only the relevant sentences, the model was trained using the embeddings created from a pretrained model called "all-MiniLM-L6-v2" [2]. An analysis of the results was made by dividing them into train and validation datasets to test the model. The results of this test will be explained in section 4. Figure 1 represent the steps followed to compute the model and how we used it.

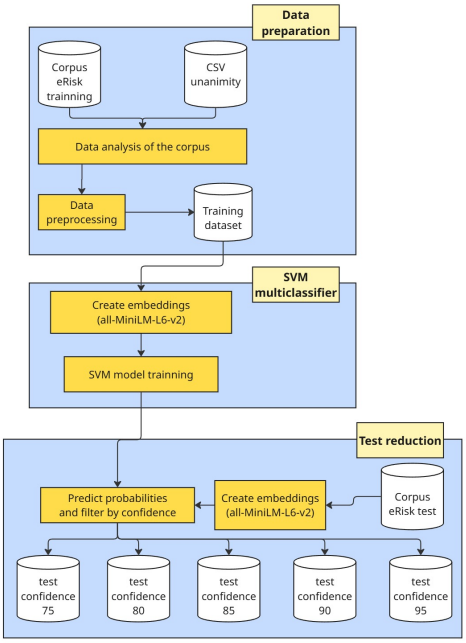

Figure 1: Architecture overview proposed for the multi classification problem

We trained the SVM with the hyperparameter probabilities set to true from the implementation given at sklearn [3] in order to predict the test sentences and filter the only ones that have higher probabilities than the threshold applied. Three different thresholds have been tested and the values are 0.75, 0.80, 0.85,0.90 and 0.95. The amount of sentences after each filter are represented in the table 1.

[1] https://scikit-learn.org/stable/modules/svm.html
[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[3] https://scikit-learn.org/stable/modules/svm.html

Table 1
Number of sentences per BDI-II symptom by confidence level

ID	Symptom (BDI-II)	75	80	85	90	95
1	Sadness	4975	3628	2436	1202	255
2	Pessimism	3350	2206	1297	608	146
3	Feelings of failure	932	596	400	239	90
4	Loss of pleasure	527	311	153	64	6
5	Feelings of guilt	1751	1436	1095	754	413
6	Feelings of punishment	1012	685	414	211	55
7	Dissatisfaction with oneself	1912	1231	753	430	207
8	Self-criticism	3141	2308	1537	813	252
9	Suicidal thoughts or wishes	4585	3437	2461	1523	641
10	Crying	7788	6259	4597	2876	1011
11	Agitation	10892	8224	5761	3435	1273
12	Loss of interest	494	290	147	60	12
13	Indecisiveness	4323	3069	1925	967	224
14	Feelings of worthlessness	263	215	159	108	56
15	Loss of energy	1861	1346	949	547	215
16	Changes in sleep	6264	5014	3725	2388	1023
17	Irritability	9928	6454	3793	1875	544
18	Changes in appetite	14478	10088	6349	3192	845
19	Difficulty concentrating	1534	1115	742	439	157
20	Fatigue or tiredness	3672	2956	2197	1436	609
21	Loss of interest in sex	3481	2144	1121	449	68

3.2. Selection of the sentences

Once we have filtered the messages from the test dataset, a sample of at most 1000 of these test messages was selected, which are the ones we have to send as runs for the proposed task. For this purpose, three different approaches have been implemented for the selection of the messages.

The first one (Figure 2) was to select the top 1000 sentences by confidence percentage, in other words, the ones that the multi-classification model gave the highest probability of belonging to that symptom. The pros and cons of this selection are clear, the pros is that the best samples will be sent to competition and therefore better results are expected, however, it is possible that it does not contain some of the symptoms as they are not classified with high percentage of confidence, and therefore, if the evaluation is an average between the results of all the symptoms, it can lead to score 0 on them if it does not contain any evaluation in this regard.

The second way to take this sample was to use the sentences previously rated above 0.95 confidence. We chose 0.95 because the other ones have a lot number of sentences for some of the symptoms and may introduce a lot of randomness to the selection of the sentences. The table 1 would be the amounts that would remain for each symptom after filtering by this confidence number.

Now to select the 1000, they will be taken proportionally to each group so that for each symptom its contribution is calculated as a percentage of 1000 and sentences are randomly selected from the subset already extracted. In case of decimals, we round down to the nearest whole number. We do this to ensure that the number of sentences does not exceed 1000. This approach ensures that you have sentences of all symptoms. Figure 3 shows the process. The following formula would be a formalization of the above applied for each symptom where Ni is the amount of sentences for the symptom and Nj is the sum of the amount of all the sentences from the 21 symptoms:

$$n_i = \left\lfloor \frac{N_i}{\sum_{j=1}^{21} N_j} \times 1000 \right\rfloor$$

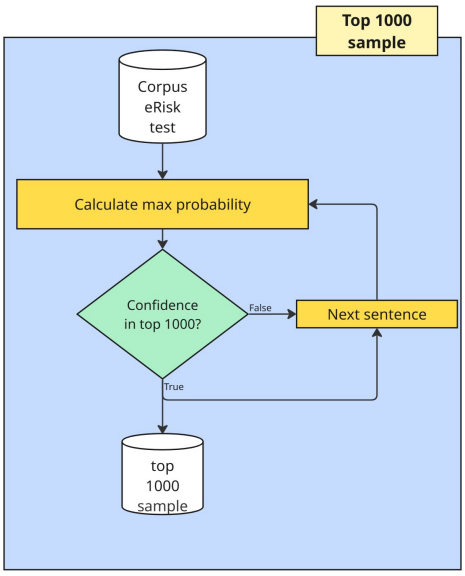

Figure 2: Process to select the top 1000 sentences

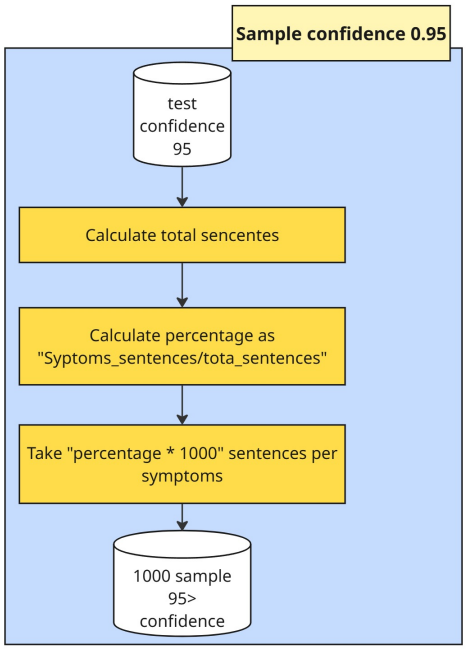

Figure 3: Process to select the sample 1000 sentences

In the last approach, after manually reviewing the sentences some of them did not talk about the symptom as if they were feeling it themselves but as if it was felt by someone else. A sub-selection of reflexive sentences was implemented within the sub dataset of 0.90 confidence (figure 4). We took a lower confidence because for some of the symptoms we did not reach the minimum we were looking for in this test. Those containing reflexive pronouns or the first person english pronouns such as 'I' or 'I'm' and their variants were selected. We then took the same number of sentences from all the symptoms to test with an equal distribution. Rounding up, this would leave 47 sentences per symptom.

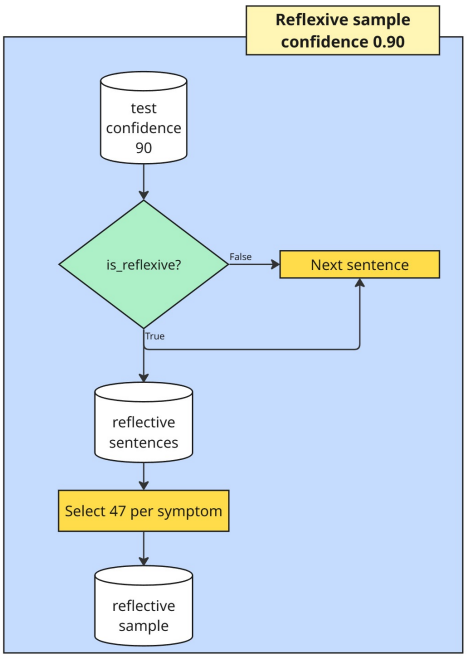

Figure 4: Process to select the sample 1000 reflexives sentences

3.3. Scoring the sentences

The next step is to score the selected sentences out of 10 and two approaches have been implemented, one using a model called VADER [4] and the other one called roberta-base-sentiment [5]. Both of them are models that are used especially for binary ranking but they have also been used for scoring in some of the cases.

In the case of VADER, it returns 4 values for each parsed sentence. Three of them are values between 0 and 1 referring to how positive, negative or neutral the sentence is, giving the sum between them a total of 1. On the other hand, the parameter 'compound' is a number between -1 and 1 that combines the three previous values, that is, the more negative the sentence, the more negative the value of the compound and vice versa.

The objective was to test what values it gave for sentences that were cataloged with some of the symptoms of depression treated in this task, and then with the use of a formula adjust that value over 10. When analyzing the sentences we saw that the sentences were never completely negative, or completely positive but especially it was more difficult for them to be cataloged as positive, so we added a multiplier to this score to increase the difference between the symptoms of greater severity with those of lesser severity, putting a multiplier of 1.1 to the negative, and 1.2 to the positive respectively, as long as the opposite was 0. That is, if a sentence had only negative and neutral values, the multiplier would be applied, but if a sentence had all 3 values, or at least positive and negative, only the 'compound' value would be taken into account. However, it should be noted that if VADER returns a positive value, it means that the phrase has no negative connotation, so it should have a low score out of 10 and the opposite is applied if VADER returns a negative value. The following formula represents how the value of 'compound', together with the multiplier explained above, was used to calculate the score rounded to two decimal places. Its important to take into account that the returned value has a maximum of 10 and a minimum of 0.

$$\left(\frac{1 - \text{sentiment\_score} \cdot \text{regulator}}{2}\right) \times 10$$

[4] https://www.nltk.org/api/nltk.sentiment.vader.html
[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

where sentiment_score is the value of compound and the regulator is the 1.1 or 1.2 value explained before.

On the other hand, the roberta-base-sentiment model is a very similar approach to VADER. In this case, the model returns a label called label and another called score. Label has 3 values, LABEL_0, LABEL_1 and LABEL_2 where 0 represents sentences categorized as negative, 1 represents neutral sentences and 2 represents positive sentences. On the other hand, the score is a value between 0 and 1 that actually refers to the confidence of assigning it to the label (positive,neutral,negative). What we have done in this case is that if it is a negative label (0), we multiply the score by 10, if it is neutral (1) we multiply it by 5 and if it is positive (2) we multiply 1-score by 10. In this way we ensure that sentences with more severe symptoms are given a higher score.

4. Results and discussion

Some internal tests were done for the multi classification task, as it is the only part we can really test it in this way due to the lack of labeled data for the scores. In that case, we divide the sentences into train and test, 80% and 20% respectively. The results show metrics as precision, recall and F1-score, including also the amount of sentences used in test with the column name of support. The results are shown in Table 2 divided by symptoms.

Table 2
Performance per symptom for the classification SVM model on a test set (Precision, Recall, F1-Score, and Support)

ID	Symptom	Precision	Recall	F1-Score	Support
1	Sadness	0.93	0.93	0.93	45
2	Pessimism	0.96	0.96	0.96	24
3	Feelings of failure	0.91	0.94	0.92	32
4	Loss of pleasure	0.91	0.91	0.91	23
5	Feelings of guilt	1.00	1.00	1.00	62
6	Feelings of punishment	1.00	1.00	1.00	17
7	Dissatisfaction with self	0.95	0.95	0.95	62
8	Self-criticism	0.97	0.95	0.96	38
9	Suicidal thoughts or desires	0.98	1.00	0.99	65
10	Crying	0.98	0.98	0.98	62
11	Agitation	0.96	1.00	0.98	55
12	Loss of interest	0.91	0.81	0.86	26
13	Indecision	0.97	0.97	0.97	33
14	Worthlessness	0.98	1.00	0.99	45
15	Loss of energy	1.00	0.94	0.97	36
16	Sleep changes	0.96	0.98	0.97	49
17	Irritability	0.97	0.92	0.95	39
18	Appetite changes	1.00	1.00	1.00	44
19	Concentration difficulty	1.00	1.00	1.00	41
20	Fatigue or tiredness	0.94	0.92	0.93	48
21	Loss of interest in sex	0.90	1.00	0.95	27

Table 3 displays the results given by the eRisk organizers for the participants for task 1, in order to compare our performance with the best team in the task. The table represents the results for majority, meaning that at least 2 of the 3 assessor marked it as correct.

Our submitted runs were a mixed of the previously explained methods, mixing different approached to take the 1000 sentences and some different ways of scoring them. There was a typo in one of the runs for roBERTa as it has the same name for two runs, but one run was using roBERTa scorer with the top 1000 sentences by confidence so should be called roBERTa top, while the other one was using roBERTa scorer again but with the sample of 1000 sentences instead of the top ones. The run starting

Table 3
Performance of HULAT-UC3M and INESC-ID for task 1 at eRisk 2025, ranked by metrics as AP, R-PREC, P@10 and NDCG with the majority results

Team	Run name	AP	R-PREC	P@10	NDCG
INESC-ID	maxcos	0.235	0.320	0.757	0.506
INESC-ID	unanimity	0.354	0.433	0.876	0.575
INESC-ID	max	0.350	0.407	0.648	0.653
INESC-ID	mix23	0.312	0.377	0.643	0.616
INESC-ID	aug-best	0.247	0.324	0.691	0.560
HULAT_UC3M	roBERTa	0.018	0.034	0.363	0.065
HULAT_UC3M	vader top	0.015	0.034	0.295	0.059
HULAT_UC3M	reflexives roBERTa	0.013	0.025	0.262	0.052
HULAT_UC3M	roBERTa	0.004	0.010	0.162	0.026
HULAT_UC3M	vader sample	0.004	0.010	0.148	0.023

with the name of VADER uses the model VADER to score the sentences, being the one called VADER top scoring the top sentences by confidence and VADER sample scoring the sample chosen sentences. Finally, relfexives roBERTa uses the reflexive sample and roBERTa scorer.

Across all evaluation metrics—Average Precision (AP), R-Precision (R-PREC), and Normalized Discounted Cumulative Gain (NDCG), our runs performed significantly worse compared to the INESC-ID team, that was the team with the best scores out of all the teams participating in the task. Their submissions achieved consistently strong results, with the best run reaching an AP score of 0.354.

Our best performing run was the one using the RoBERTa model, where we selected the top 1000 sentences based on confidence scores from our multiclass classification model with an AP of 0.018. This was closely followed by a similar run using the VADER model for scoring with the same top 1000 sentence selection approach. These two runs outperformed our other three submission by a wide margin in fact, they were up to four times more precise than the runs that used sentence sampling (runs 4 and 5) with a confidence threshold of 0.95, as previously described in this document.

From these results, we can conclude that the multi class classification model was effective. The two runs that used high confidence sentence selection clearly outperformed the runs based on lower confidence sampling, suggesting that confidence based filtering had a strong positive impact on performance. However, the methods used to score the sentences are not appropriate.

Table 4 shows the result in case of unanimity, meaning that all the three annotators have to agree on the sentence being well classified and scored.

Table 4
Performance of HULAT UC3M and INESC-ID for task 1 at eRisk 2025, ranked by metrics as AP, R-PREC, P@10 and NDCG with the majority results

Team	Run name	AP	R-PREC	P@10	NDCG
INESC-ID	unanimity	0.269	0.383	0.509	0.561
INESC-ID	max	0.223	0.308	0.386	0.582
INESC-ID	mix23	0.201	0.279	0.371	0.547
INESC-ID	aug-best	0.167	0.236	0.414	0.515
INESC-ID	maxcos	0.164	0.273	0.429	0.472
HULAT_UC3M	reflexives roBERTa	0.013	0.032	0.157	0.053
HULAT_UC3M	roBERTa	0.008	0.025	0.174	0.040
HULAT_UC3M	vader top	0.006	0.024	0.105	0.037
HULAT_UC3M	roBERTa	0.002	0.009	0.052	0.016
HULAT_UC3M	vader sample	0.001	0.009	0.029	0.012

In this case, the best run is the one that chooses reflexive sentences. It does not get worse compared to the results given in the table 3 representing the result of majority, 0.013 of precision in both cases. This result can lead us to think that if a sentence is reflexive, is more likely to be correctly selected than if it is not, as all the sentences correctly scored were all ranked by unanimity as in majority we have the same score compared to unanimity. In the case of the other runs, it gets worse by half or more of the precision for our team, while the score from the best team in this case is the run called unanimity, which can suggest that they used the data labeled as unanimity from the corpus provided by eRisk. The result achieved from INESC-ID can lead to think that our proposal of using only the unanimity sentences to train the classifier was in the good way, removing part of the noise or not clear sentences from the labeled data.

5. Conclusions and Future Work

We have presented a general overview of our participation at eRisk task 1, search for symptoms of depression using mainly machine learning approaches. As we mentioned in the previous section, we can notice that the approaches were not good in general even though we could get some possible conclusions or future testing.

For possible future work, we would like to change the way of scoring the sentences, as we could notice thanks to some conclusions that it was the weakest point of our systems. Probably VADER and roBERTa models are not adequate for this task, or we have not fixed fine tune it enough in order to get the best use of it. For substituting this approaches we would like to try generative AI for generating labeled data with the score giving in the prompt some examples of the labeled data and the BDI-II Sen. Some other teams, as mentioned in the related work section used it to generate data in general, but we would like to use it only for creating data with the scores with the idea of using it directly in a machine learning architecture as a SVM, using a two level architecture, firstly the SVM classifier followed by the SVM regressor for each of the symptoms for creating the scores.

Other solution could be to explore generative AI directly to score the sentences previously chose without training a machine learning architecture nor generating new data using generative AI. In this case, the scores will directly depend on the prompt used to generate the data so it will be important to give several accurate examples to the generative AI in order to achieve good results. In this approach we would like to have a professional in depression for creating some scored sentences for each of the symptoms to introduce them in the prompt. This approach without good and accurate examples do not work as it depends directly in the random choices made by the AI.

In other hand, some deep learning approaches would be really helpful for generating the scores. One of the approaches that would like test is training a neural network (NN) with very extreme sentences of depressed users in the internet and use a softmax classifier at the last layers of that NN. The purpose of the approach is to return the probability of the sentence to be part of that symptom as a list of 21 values, one for each of the symptoms. Then the score will be computed by multiplying the maximum probability between all of the symptoms by 10 in order to get the score and we will assign the sentence to the symptom having the higher probability. In the end, the sentences clearly being part of one symptom will get very high score and, if the model is unclear about it, the probability will be very low, and therefore the score will follow to as depends on the probabilities returned by the model. This approach substitutes the previously mentioned multiclassifier and requires appropriate data, as the labeled data given by the eRisk task do not have only severe sentences for each symptom, so we will need to combine data generation with generative AI or similar approaches for doing data augmentation in combination with this approach.

Acknowledgments

This work was partially supported by Grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN AI) by MICIU/AEI/ 10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

The team has used generative AI, particularly ChatGPT, in spelling check for creating this document and code regarding latex format. In addition some minor errors in the code has been fixed using it.

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[2] W. health organization (WHO), Depressive disorder (depression), 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression, last access: 16 de mayo de 2025.

[3] W. health organization (WHO), Depressive disorder (depression), 2025. URL: https://www.who.int/news-room/fact-sheets/detail/suicide, last access: 16 de mayo de 2025.

[4] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[5] S. G. Burdisso, M. Errecalde, M. M. y Gómez, Using text classification to estimate the depression level of reddit users, Journal of Computer Science and Technology, Vol 21, Iss 1, Pp e1-e1 (2021) (2021). URL: https://doi.org/10.24215/16666038.21.e1. doi:10.24215/16666038.21.e1.

[6] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.

[7] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, Springer, 2019, pp. 340–357.

[8] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020 (2020).

[9] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) 1 (2021) 864–887.

[10] S.-H. Wu, Z.-J. Qiu, A roberta-based model on measuring the severity of the signs of depression., in: CLEF (Working Notes), 2021, pp. 1071–1080.

[11] R. Manna, J. Monti, et al., Unior nlp at erisk 2021: Assessing the severity of depression with part of speech and syntactic features, in: CEUR WORKSHOP PROCEEDINGS, volume 2936, CEUR, 2021, pp. 1022–1030.

[12] L. Barros, A. Trifan, J. L. Oliveira, Vader meets bert: sentiment analysis for early detection of signs of self-harm through social mining., in: CLEF (working notes), 2021, pp. 897–907.

[13] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2022: Early risk prediction on the internet (extended overview), in: CEUR Workshop Proceedings (CEUR-WS. org), 2022.

[14] A.-M. Bucur, A. Cosma, L. P. Dinu, P. Rosso, An end-to-end set transformer for user-level classification of depression and gambling disorder, arXiv preprint arXiv:2207.00753 (2022).

[15] H. Srivastava, N. Lijin, S. Sruthi, T. Basu, Nlp-iiserb@ erisk2022: Exploring the potential of bag of words, document embeddings and transformer based framework for early prediction of eating disorder, depression and pathological gambling over social media., in: CLEF (Working Notes), 2022, pp. 972–986.

[16] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction

on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 294–315.
[17] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of general psychiatry 4 (1961) 561–571.
[18] A.-M. Bucur, Utilizing chatgpt generated data to retrieve depression symptoms from social media, arXiv preprint arXiv:2307.02313 (2023).
[19] S. Ubani, S. O. Polat, R. Nielsen, Zeroshotdataaug: Generating and augmenting training data with chatgpt, arXiv preprint arXiv:2304.14334 (2023).
[20] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, F. Zeng, W. Liu, et al., Auggpt: Leveraging chatgpt for text data augmentation, IEEE Transactions on Big Data (2025).
[21] D. Maupomé, T. Soulas, F. Rancourt, G. Cantin-Savoie, G. Winterstein, S. Mosser, M.-J. Meurs, Lightweight methods for early risk detection., in: CLEF (Working Notes), 2023, pp. 718–726.
[22] N. Recharla, P. Bolimera, Y. Gupta, A. K. Madasamy, Exploring depression symptoms through similarity methods in social media posts., in: CLEF (Working Notes), 2023, pp. 763–772.