

REBECCA at eRisk 2024: Search for Symptoms of Depression Using Sentence Embeddings and Prompt-Based Filtering

Notebook for the eRisk Lab at CLEF 2024

Anna Barachanou^{1,*}, Filareti Tsalakanidou¹ and Symeon Papadopoulos¹

¹Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece

Abstract

Depression is a complex mental health disorder characterized by persistent feelings of sadness, hopelessness, and a lack of interest or pleasure in daily activities. It significantly affects an individual's well-being, impairing their ability to work, socialize with others and be creative. Social media is used by billions of people globally who interact and generate an abundance of posts and texts. Analysis of social interaction data offers opportunities to gain valuable insights into people's mental health and potentially take supportive action. eRisk 2024 focuses on the challenge of early risk detection on the Internet and has established a number of tasks for this reason. We participated in *Task 1: Search for symptoms of depression*. The aim of this task is to rank user sentences in terms of 21 symptoms of depression. This paper presents our approach combining ranking sentences using cosine similarity and Transformer embeddings and refining our results with the use of a Large Language Model (LLM). Our LLM-refined approach was among the best performing ones among the 29 runs submitted by the 9 participating teams.

Keywords

early risk detection, natural language processing, depression, text retrieval, prompt engineering, transformers

1. Introduction

Depression is a debilitating mental health condition affecting 5% of people worldwide according to WHO (World Health Organization)¹. Individuals suffering from depression experience a variety of symptoms beyond a persistently depressed mood and dysphoria. Depression may also manifest as a loss of interest in activities they once enjoyed, significant changes in sleep and appetite, feelings of guilt and hopelessness, fatigue, restlessness, problems with concentration and even suicidal ideation [1]. Beck's Depression Inventory (BDI-II) [2] is one of the most widely used psychometric assessment tools for depression and it is designed in the form of a questionnaire measuring the severity of such symptoms of depression in adolescents and adults.

In today's digitally connected world, social media such as Facebook, Instagram, YouTube, Twitter, etc. are being used by more than 4.76 billion people worldwide². Among these users, there are many people affected by mental health conditions including depression. Through social media, users interact and share their thoughts, opinions and emotions with others. As a result, there are vast amounts of data generated every day that could potentially be leveraged to provide insights into their mental well-being. This presents a unique opportunity for mental health professionals and researchers to analyze language patterns by using modern Natural Language Processing (NLP) techniques. By examining the textual content shared on social media, it should be possible to build methods for early detection of depression.

Early detection of risk factors such as depression can prevent numerous negative outcomes to an individual's life. Recognizing and addressing symptoms of depression early on, facilitates timely helpful

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

✉ barachanou@iti.gr (A. Barachanou); filareti@iti.gr (F. Tsalakanidou); papadop@iti.gr (S. Papadopoulos)

🆔 0009-0007-1193-7682 (A. Barachanou); 0000-0002-5310-8045 (F. Tsalakanidou); 0000-0002-5441-7341 (S. Papadopoulos)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.who.int/news-room/fact-sheets/detail/depression>

²<https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>

intervention and support, which can significantly improve the effectiveness of treatment. Individuals are more likely to respond positively to treatment when intervention begins early and avoid intensified and persistent symptoms. Especially because depression is a major risk factor for suicide, the early offer of support can potentially minimize the risk of suicide and suicidal behaviors. This will eventually lead to an enhanced quality of life free of symptoms of depression, enabling individuals to engage actively and socially in their everyday life.

The eRisk lab of CLEF (Conference and Labs of the Evaluation Forum), focuses on early risk prediction on the Internet. Ever since their beginning in 2017 when they piloted a task on Early Detection of depression [3], eRisk’s primary objective has been depression which then expanded to include tasks related to other mental illnesses as well. In this paper, we present our participation, motivated by our involvement in the Horizon 2020 REBECCA project, in Task 1: Search for symptoms of depression. This task is a continuation of the same task in eRisk 2023. We were inspired by the systems developed by the participating teams and we attempted to improve results with the use of Large Language Models (LLMs) and prompt engineering. Moreover, traditional information retrieval methods such as BM25 or TF-IDF can effectively handle document ranking but often lack the semantic depth needed for precise results. Given that depression is a complex and delicate subject, there is a need for highly accurate methods in order to rank sentences with respect to depression symptoms. We initially ranked the sentences to the symptoms using Transformer embeddings and computed the ranking scores with cosine similarity. Subsequently, we decided to leverage the reasoning capabilities of an LLM (namely GPT-4) to refine the results of the base method, emulating the process of providing relevance feedback and removing all non-relevant sentences that do not reflect the author’s state about the symptoms. Our methods achieved highly competitive results among the 9 participating teams, outperforming all competing approaches in terms of Precision@10 in the unanimity setting, and revealed potential in the systems we developed and especially in utilizing GPT-4 to better grasp the concepts of depression in sentences.

2. Related work

A significant portion of the related literature about depression is focused on depression identification. For example, Jamil et al. [4] aimed to identify depression from individual tweets and assess the risk of depression from a user’s set of tweets. They computed a small number of features, using indicators like the percentage of depressed tweets, self-reported depression, BOW, and other lexical features. They also employed SVM for classification and used balancing methods like undersampling and SMOTE. Similarly, Peng et al. [5] used various ML models and multi-kernel SVM to combine features from a user’s texts, profile and behaviour. While, Chen et al. [6] used emotion analysis with EMOTIVE [7], linguistic features from LIWC, and behavioral features to identify mental health conditions and employed several ML models for the classification task.

eRisk 2023 [8] established three tasks surrounding mental health, including Task 1: Search for symptoms of depression. The task we are currently participating in is a continuation of this, with the aim of expanding research further on this promising topic. The Formula-ML team [9] achieved the best performance by leveraging Transformer embeddings and word2vec for sentence embeddings. Thereafter, they applied soft cosine similarity between sentences and BDI-II terms for each symptom and performed weighted aggregation of these scores to compute the final scores and rank the sentences in relation to symptoms of depression. A number of participating teams utilized LLMs in their systems for eRisk 2023 in various tasks. For Task 1 in particular, the BLUE team [10] utilized ChatGPT to enrich the BDI-II questionnaire terms, enhancing diversity. Then computed embeddings using two Transformer models and performed semantic similarity with cosine similarity to ultimately rank the sentences.

Large Language Models are a relatively recent innovation in the field of Artificial Intelligence (AI) and NLP; however, they already show great potential in many fields including mental health. Bakir Hadzic et al. [11] compared the efficacy of three popular LLMs: BERT, GPT-3.5 and GPT-4 for early detection depression in textual data. The research was conducted across three datasets and revealed that GPT-4

significantly outperforms both BERT and GPT-3.5, demonstrating superior performance without prior fine-tuning. This suggests that GPT-4 could be a highly effective tool for early depression detection. The study also highlights the potential of models like GPT-4 in mental health beyond depression, proposing further development and fine-tuning of LLMs.

3. Methodology

We participated in Task 1: Search for symptoms of depression for the eRisk 2024 [12] [13] lab of CLEF 2024. This is a continuation of the same task from CLEF eRisk 2023. The task consists of ranking sentences from social media in terms of 21 symptoms of depression (Table 1) from the Beck Depression Inventory–II (BDI-II) questionnaire. The BDI-II questionnaire is a self-report rating inventory and it consists of 21 multiple-choice questions, each one relating to a specific symptom. Each question has four possible answers from least to most severe, associated with a score from 0 to 3 respectively. The scores assigned to each question are then summed to a total score with a maximum score of 63. High total scores indicate a high chance of depressive symptoms.

Table 1
The 21 symptoms of depression according to BDI-II

Symptoms		
Sadness	Pessimism	Past Failure
Loss of Pleasure	Guilty Feelings	Punishment Feelings
Self-Dislike	Self-Criticalness	Suicidal Thoughts or Wishes
Crying	Agitation	Loss of Interest
Indecisiveness	Worthlessness	Loss of Energy
Changes in Sleep Patterns	Irritability	Changes in Appetite
Concentration Difficulty	Tiredness or Fatigue	Loss of Interest in Sex

In more detail, each social media sentence should be assigned to the most relevant symptom out of the 21. Subsequently, the sentences assigned to each symptom should be ordered in decreasing order from the most to the least relevant. The relevant sentences should convey the author’s state concerning the symptom, even if the sentiment is positive. For example, a sentence that expresses happiness should be also considered relevant to the symptom of sadness. It is also emphasized that a sentence is only relevant when it is solely about the author’s feelings related to the symptom and not the feelings of other individuals. For example, a user post mentioning that the user’s sister is sad is not considered relevant to sadness for that user because the user is not sad but their sister is.

3.1. Dataset

We were provided with two TREC formatted sentence-tagged datasets, one for training and one for testing. Both datasets consist of unlabeled user sentences from Reddit posts. The training dataset consists of last year’s data and the test set contains new data for this year’s eRisk that are to be used for the evaluation of our systems. As presented in Table 2, the test data consist of a total of 15M sentences, which is 11M more sentences than the dataset used in 2023, with approximately 18 words in a sentence on average. We additionally created a small third dataset containing all symptoms and their respective relevant answers from the BDI-II questionnaire. Examples for the symptoms of sadness and pessimism are presented in Table 3.

3.2. Ranking system

The system we developed is illustrated in the flowchart of Figure 1. It involves multiple steps that we will expand on below. These include text pre-processing, dataset cleaning by discarding sentences that

Table 2
Corpus statistics

	Training	Test
Number of sentences	4M	15M
Number of users	3,106	553
Avg number of words/sentence	13.99	17.99

Table 3
Relevant answers from the BDI-II for sadness and pessimism

Sadness	Pessimism
I do not feel sad	I am not discouraged about my future
I feel sad much of the time	I feel more discouraged about my future than I used to
I am sad all the time	I do not expect things to work out for me
I am so sad or unhappy that I can't stand it	I feel my future is hopeless and will only get worse

are not about the authors, sentence ranking using a pre-trained Transformer for sentence embeddings and cosine similarity, and result refinement using GPT-4.

For pre-processing, we translated all texts to English, turned all texts into lowercase, removed punctuation and non-alphabetic symbols, and fixed word contractions. A sentence is considered relevant only when it reflects the author's state surrounding a symptom, consequently we conducted keyword matching in order to only keep sentences that indicate that the author is talking about themselves (I, me, mine, myself, mine, we, us, our, ourselves, ours). Following the removal of sentences not containing any of the aforementioned keywords, we are confident that we have eliminated a substantial portion of irrelevant texts, simultaneously reducing the computational workload from 15M to 11M sentences (Table 4).

Table 4
Number of sentences after cleaning

	Training	Test
Initial number of sentences	4M	15M
Number of sentences after elimination	1M	11M

Due to the datasets provided being unlabeled, we focused on unsupervised methods for our systems. We chose a pre-trained Transformer model to calculate the embeddings for the sentences and the answers of each symptom. The Massive Text Embedding Benchmark (MTEB) [14] evaluates text embeddings across a broad range of tasks and datasets to provide a comprehensive assessment of their performance. It spans 8 embedding tasks, 58 datasets, and 112 languages and tests models to determine their effectiveness. The MTEB Leaderboard³ presents all tested models across all tasks, including text ranking, along with numerous evaluation metrics. We considered models for the Retrieval and Reranking tasks that were evaluated using NDCG@ k (Normalized Discounted Gain at k) and MAP (Mean Average Precision), respectively. Since we could already expect how some Transformer-based embeddings would perform thanks to last year's submissions, we explored new Transformer models for this part of the task, by excluding models that were involved in last year's submissions. Based on the above criteria and the need for a model that is as lightweight as possible without sacrificing substantial performance, we opted for the `bge-small-en-v1.5`⁴ Transformer model [15] that calculates 384-dimensional embeddings and consists of 33M parameters.

We calculated the cosine similarity score for each sentence paired with each answer for every

³<https://huggingface.co/spaces/mteb/leaderboard>

⁴<https://huggingface.co/BAAI/bge-small-en-v1.5>

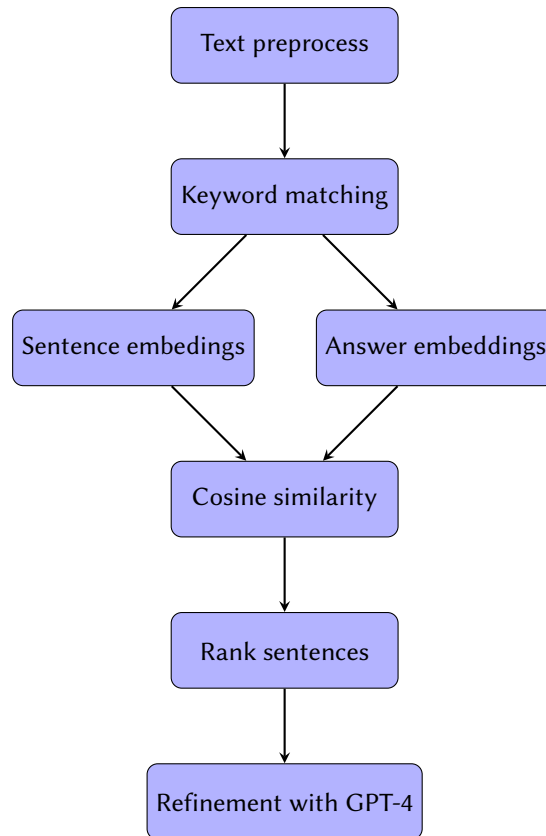


Figure 1: Methodology flowchart

symptom. For every sentence, we kept the max similarity score out of the sentence-answer pairs for each symptom and then assigned the sentence to the symptom with total max score. We then ranked the sentences under every symptom based on the above score and kept the top 1,000 per symptom resulting in a total of 21,000 ranked sentences from the initial corpus.

Depression is a complex and delicate subject, hence we expect that our initial ranking using the above method would be a decent but crude approximation to the task. To further refine our results, we resorted to prompt engineering on top of GPT-4, which is considered as one of the state-of-the-art LLMs. We used prompt engineering to discard any non-relevant sentences that were ranked high by the previous steps of our system. Our goal was to use prompts asking GPT-4 to decide whether a sentence is actually relevant (according to GPT-4) to the symptom. We first conducted experiments using ChatGPT testing various candidate prompts comparing a shared prompt strategy (i.e. using the same prompt for all symptoms) versus a symptom-specific prompt strategy. After our initial experimentation, we decided that a symptom-specific strategy was more effective. All symptom-specific prompts followed the same syntax for the sake of uniformity. Subsequently, we used the more powerful gpt4-turbo model that we accessed via the OpenAI API⁵ for the final results. Our 21 prompts followed the subsequent structure: “We will provide you with some sentences. Your task is to decide whether they are related to symptom in a positive/negative sentiment or not”. Where we included each symptom and its respective positive and negative sentiment. The detailed prompts are provided in Appendix A.

Since positive feelings about a symptom are to be considered relevant as well, we made an effort to include positive sentiment in the relevant sentences using our prompt. We removed all text that was not considered relevant by GPT-4 resulting in 14,815 sentences, meaning that 6,185 sentences were discarded as non-relevant. We submitted both the method without prompt engineering and with GPT-4 assessment in order to evaluate if GPT-4 improved the overall performance.

⁵<https://openai.com/api/>

4. Results

We provided two runs with our results in the requested TREC format: **TransformerEmbeddings_CosineSimilarity** containing the results of our baseline method and **TransformerEmbeddings_CosineSimilarity_gpt** with our final results using ranking refinement with GPT-4. In total, 9 teams participated in eRisk 2024 Task 1 with 29 submitted runs.

eRisk selected a number of sentences from all teams’ submissions using top-k pooling. Then the assessment was performed by human assessors who examined whether a sentence was correctly ranked to a symptom or not. Two types of evaluations took place: a) a *majority vote* where the agreement of the majority of the assessors is enough to label a ranking as correct (or not); b) a *unanimity vote* where all of the assessors are required to agree. Five metrics were used for the evaluation of all submissions: AP (Average Precision), MAP (Mean Average Precision), R-PREC (Recall Precision), P@10 (Precision at 10) and NDCG (Normalized Discounted Cumulative Gain).

As presented in Tables 5 and 6, our systems demonstrated good performance across all metrics in both the majority and the unanimity vote. Regarding the majority vote, we are approaching the performance levels of the top performing teams across all metrics and we are above the mean and median of total runs submitted by all teams. While our method with GPT-4 ranking refinement **TransformerEmbeddings_CosineSimilarity_gpt** is improving the performance across all scores except from NDCG.

Table 5
Majority voting results

Team	Method	AP	R-PREC	P10	NDCG
MeVer-REBECCA	TransformerEmbeddings_CosineSimilarity_gpt	0.301	0.340	0.981	0.506
MeVer-REBECCA	TransformerEmbeddings_CosineSimilarity	0.295	0.332	0.976	0.517
NUS-IDS	Config 5	0.375	0.434	0.924	0.631
APB-UC3M	APB-UC3M sentsim-all-MiniLM-L6-v2	0.354	0.391	0.986	0.591
All team runs	Mean	0.226	0.253	0.685	0.375
All team runs	Median	0.252	0.322	0.738	0.453

Concerning the unanimity vote, we received the best P@10 score of 0.833 for **TransformerEmbeddings_CosineSimilarity_gpt** amongst all 29 runs of the participating teams. In terms of the rest of the metrics, we are close to the best performing team, while our scores exceed both the mean and median values of the scores of all teams runs once again. Consequently, the results indicate the strength of both our baseline model and our refinement method. Our ranking refinement proposal turned out to improve overall performance as there was an increase across all metrics with the exception of NDCG.

Table 6
Unanimity voting results

Team	Method	AP	R-PREC	P10	NDCG
MeVer-REBECCA	TransformerEmbeddings_CosineSimilarity_gpt	0.305	0.357	0.833	0.551
MeVer-REBECCA	TransformerEmbeddings_CosineSimilarity	0.294	0.349	0.824	0.556
NUS-IDS	Config 5	0.392	0.436	0.795	0.692
All team runs	Mean	0.220	0.248	0.548	0.411
All team runs	Median	0.227	0.275	0.576	0.499

5. Conclusion and future work

In conclusion, based on the mean and median of the assessment scores of all teams, our methods are competitive and exhibit potential for future research. Our proposed methodology consisted of a few

pre-processing and cleaning steps followed by a simple ranking using sentence embeddings, which was further refined based on a prompt engineering strategy on top of GPT-4. However, there is room for improvement in the scores by making enhancements in our methodology. One future step is to experiment with various other prompting strategies that could be more effective in detecting relevant and non-relevant sentences. Moreover, one could leverage publicly available depression-annotated corpora to fine-tune GPT-4 so that it can better recognize the relevance of sentences to depression symptoms. Finally, we could investigate leveraging LLMs to annotate parts of the dataset and use these to train more accurate deep learning models in a supervised manner.

Acknowledgments

This work has been partially funded by the H2020 project “REBECCA: REsearch on BrEast Cancer induced chronic conditions supported by Causal Analysis of multi-source data” under Grant Agreement no. 965231 (<https://rebeccaproject.eu/>).

References

- [1] J. W. Kanter, A. M. Busch, C. E. Weeks, S. J. Landes, The nature of clinical depression: symptoms, syndromes, and behavior analysis, *The Behavior analyst* 31 (2008) 1–21. doi:10.1007/BF03392158.
- [2] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, *JAMA Psychiatry* 4 (1961) 561–571. doi:10.1001/archpsyc.1961.01710120031004.
- [3] D. Losada, F. Crestani, A test collection for research on depression and language use, volume 9822, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.
- [4] Z. Jamil, D. Inkpen, P. Buddhitha, K. White, Monitoring tweets for depression to detect at-risk users, in: K. Hollingshead, M. E. Ireland, K. Loveys (Eds.), *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality*, Association for Computational Linguistics, Vancouver, BC, 2017, pp. 32–40. URL: <https://aclanthology.org/W17-3104>. doi:10.18653/v1/W17-3104.
- [5] Z. Peng, Q. Hu, J. Dang, Multi-kernel svm based depression recognition using social media data, *International Journal of Machine Learning and Cybernetics* 10 (2017) 43 – 57. doi:10.1007/s13042-017-0697-1.
- [6] X. Chen, M. Sykora, T. Jackson, S. Elayan, F. Munir, Tweeting your mental health: an exploration of different classifiers and features with emotional signals in identifying mental health conditions, 2018. doi:10.24251/HICSS.2018.421.
- [7] M. Sykora, T. Jackson, A. O’Brien, S. Elayan, Emotive ontology: Extracting fine-grained emotions from terse, informal messages, *International Journal on Computer Science and Information Systems* 8 (2013) 106–118.
- [8] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, Springer-Verlag, Berlin, Heidelberg, 2023, p. 294–315. URL: https://doi.org/10.1007/978-3-031-42448-9_22. doi:10.1007/978-3-031-42448-9_22.
- [9] N. Recharla, P. Bolimera, Y. Gupta, A. K. Madasamy, Exploring depression symptoms through similarity methods in social media posts, 2023. URL: <https://ceur-ws.org/Vol-3497/paper-065.pdf>.
- [10] A.-M. Bucur, Utilizing chatgpt generated data to retrieve depression symptoms from social media, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, 2023, pp. 662–671. URL: <https://ceur-ws.org/Vol-3497/paper-055.pdf>.
- [11] B. Hadzic, P. Mohammed, M. Danner, J. Ohse, Y. Zhang, Y. Shiban, M. Rättsch, Enhancing early depression detection with ai: a comparative use of nlp models, *SICE Journal of Control, Measurement, and System Integration* 17 (2024) 135–143. doi:10.1080/18824889.2024.2342624.

- [12] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024, Springer International, Grenoble, France, 2024.
- [13] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, Grenoble, France, September 9th to 12th, 2024, CLEF 2024, CEUR Workshop Proceedings, 2024.
- [14] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive text embedding benchmark, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2014–2037. doi:10.18653/v1/2023.eacl-main.148.
- [15] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.

A. GPT-4 Prompts

Here we present the prompts used in this work in detail:

1. We will provide you with some sentences. Your task is to decide whether they are related to sadness/happiness or not.
2. We will provide you with some sentences. Your task is to decide whether they are relevant to pessimism/optimism or not.
3. We will provide you with some sentences. Your task is to decide whether they are relevant to past failure/success or not.
4. We will provide you with some sentences. Your task is to decide whether they are relevant to the recent loss (or not) of pleasure.
5. We will provide you with some sentences. Your task is to decide whether they are relevant to feeling (or not feeling) guilty.
6. We will provide you with some sentences. Your task is to decide whether they are relevant to someone feeling like they themselves are being (or not being) punished.
7. We will provide you with some sentences. Your task is to decide whether they are relevant to someone disliking or liking themselves.
8. We will provide you with some sentences. Your task is to decide whether they are relevant to someone feeling (or not feeling) critical towards themselves.
9. We will provide you with some sentences. Your task is to decide whether they are relevant to having (or not having) suicidal thoughts and wishes.
10. We will provide you with some sentences. Your task is to decide whether the sentences mention crying or not crying now or any other time.
11. We will provide you with some sentences. Your task is to decide whether they are relevant to feeling (or not feeling) agitated.
12. We will provide you with some sentences. Your task is to decide whether they are relevant to losing (or not losing) interest in things.
13. We will provide you with some sentences. Your task is to decide whether they are relevant to being (or not being) indecisive.
14. We will provide you with some sentences. Your task is to decide whether they are relevant to feeling (or not feeling) worthless.
15. We will provide you with some sentences. Your task is to decide whether they are relevant to having (or not having) energy.

16. We will provide you with some sentences. Your task is to decide whether they are relevant to experiencing (or not experiencing) changes in sleeping pattern.
17. We will provide you with some sentences. Your task is to decide whether they are relevant to feeling (or not feeling) irritable.
18. We will provide you with some sentences. Your task is to decide whether they are relevant to experiencing (or not experiencing) changes in appetite.
19. We will provide you with some sentences. Your task is to decide whether they are relevant to having (or not having) difficulty concentrating.
20. We will provide you with some sentences. Your task is to decide whether they are relevant to feeling (or not feeling) tired.
21. We will provide you with some sentences. Your task is to decide whether they are relevant to losing (or not losing) interest in sex.