THINKIR at eRisk 2025: Early Detection and Risk Assessment of Depression using Transformer Models

Notebook for the eRisk Lab at CLEF 2025

Subinay Adhikary[1,*], Junsume Das[2] and Dwaipayan Roy[1]

[1]Indian Institute of Science Education and Research Kolkata, Department of Computational and Data Sciences
[2]Vellore Institute of Technology, Department of Mathematics, India

Abstract

Depression is one of the leading causes of disability worldwide, affecting 5% of adults, as many persons are not aware of the symptoms of mental disorders. Due to the tendency of people to share their thoughts on social platforms such as Reddit, to identify user's mental states social media information can be utilized, to measure depression levels some metrics are also available such as Beck's Depression Inventory-II (BDI-II). BDI-II consists of 21 symptoms of depression, which cover sadness, failure, guilt, etc, and we aim to retrieve relevant documents for each symptom from the provided corpus. In this work, we (team THINKIR) showcase that pseudo-relevant documents aid in enlisting the new terms in the existing query set. This year, our group THINKIR takes part in **Task 1** of eRisk 2025, which focuses on ranking sentences from a set according to their relevance to the BDI-II symptoms. We concentrate on creating strategies that enhance the accuracy and scope of symptom-specific relevance ranking, utilizing both query-driven retrieval techniques and classification models aware of symptoms. Through utilizing diverse but complementary methods, we seek to improve the comprehension and recognition of depression-related phrases in user-generated content. Motivated by this, our team THINKIR investigates various strategies from query-driven retrieval to classification modelsâĂŤto efficiently pinpoint symptom-related content in the given corpus.

Keywords

query expansion, depression detection, large language model

1. Introduction

Broadly human has two different kinds of emotions, including positive and negative emotions. Positive emotions include love, laughter, and happiness. On the other side, anger, sadness, depression indicate negative emotions and can lead to deaths [1]. Additionally, depression affects mental and physical health, particularly teens and younger individuals [2]. Suicide is a major and probable inevitable public problem, where depression enhances the chance of suicide [3]. Since lack of access to medical services or fear around mental illness, effective treatment for depression remains undiagnosed [4]. As a result of that depression has become the fourth leading cause of death among those aged 15-29 [5].

Mental health conditions such as depression develop gradually over time and can be identified in the early stage by measuring the symptoms using different scales. To measure depression levels, mental health professionals use other scales, such as the Center of Epidemiological Scales-depression (CES-D) [6], Patient Health Questionnarie (HRSD) [7], and Beck's Depression Inventory-II (BDI-II) [8], Hamilton Rating Scale for Depression (HRSD) [9], and measuring someone's mental health based on their action is a sophisticated psychological science that has not been explored well. It motivates researchers to focus on the identification of mental depression using social media data, specifically, NLP researchers have contributed more by using Convolutional Neural Networks (CNN) [10], Recurrent Neural Networks (RNN) [11], Hierarchical Attention Networks (HAN) [12], and transformer-based architectures [13] for mental disorder detection in last few years.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
sa21rs094@iiserkol.ac.in (S. Adhikary); junsume.das03@gmail.com (J. Das); dwaipayan.roy@iiserkol.ac.in (D. Roy)
https://dwaipayanroy.github.io/ (D. Roy)

This paper presents our approach for retrieving top-1000 sentences for each of the 21 questions from Beck's Depression Inventory-II (BDI-II), Task 1 of the eRisk 2025 shared task at CLEF 2025 [14, 15]. The main challenging factor is making a set of queries that can retrieve more relevant sentences from the collection, which leads to *query expansion problem*. With this motivation, we focus on selecting the terms from the relevant documents and adding them to the existing query set – commonly known as *pseudo relevance feedback*.

Apart from the classical query expansion method, we explore by incorporating two novel directions aimed at improving the relevance and coverage of retrieved sentences for the 21 BDI-II symptoms. First, we fine-tuned transformer-based models such as BERT [16] to learn mappings between sentence-level textual patterns and the associated BDI-II labels and where each sentence is associated with one or more depressive symptoms, which is a *multi-label text classification* task. This model-driven approach offers a complementary perspective to the traditional query-based retrieval, as it allows the system to make predictions directly based on learned representations rather than relying solely on term matching. Secondly, we leverage large language models (LLMs), motivated by recent advancements in in-context learning [17, 18, 19] and prompt engineering. In particular, for each symptom, we collect a representative set of user-generated sentences and prompt a GPT-based model to produce a refined, coherent query that captures the essence of that symptom, aka few-shot learning. These queries generated by LLMs seek to enhance retrieval accuracy by more effectively matching the search intent with the semantic context of the related symptom descriptions [18].

These two complementary approaches, classification-driven symptom detection and LLM-guided query generation strategy space for tackling symptom relevance ranking and offer promising alternatives to traditional expansion-based methods. Finally, we showcase that transformer-based model yields better performance among all the methods, as described in Section 6.

2. Related Work

Several prior works have explored Task 1 of the eRisk challenge, focusing on retrieving relevant sentences for depression symptoms. Recent contributions at CLEF 2024 include diverse approaches such as sentence embedding and prompt-based filtering [20], ensemble learning with early maladaptive schemas [21], and transformer-based classification [22]. Systems like SINAI [23], MindwaveML [24], and APB-UC3M [25] have demonstrated the effectiveness of natural language processing and deep learning for detecting depressive symptoms. These efforts highlight the growing interest and variety of strategies in addressing sentence-level symptom detection using social media data.

The approach involves authors integrating psycholinguistic and behavioral characteristics to link users' posts with their BDI responses, as detailed in [26]. Spartalis et al. [27] employed three techniques to automatically complete the BDI questionnaire. The first two approaches classified each user's choices for each BDI item using cosine similarity and well-known classifiers like SVM and Random Forest. They utilized a language model called SBERT to represent the subjects' posts. The third technique enhanced a RoBERTa model to predict the respondents' responses for the collection.

Recently, the CLEF 2024 eRisk overview document [28] addressed the wider issues of early forecasting risks in mental health through social media information. This paper provided key insights into task formulation and evaluation, which inspired the formulation of our own symptom detection task as a sentence-level relevance ranking problem grounded in textual data.

The CLEF eRisk 2024 overview [28] expanded the scope of early risk detection, emphasizing sentence-level relevance ranking for depressive symptoms. This task formulation inspired our approach to symptom detection as a sentence-level relevance ranking problem grounded in textual data.

The DepreSym dataset introduced by PÃĺrez et al. [29] provided a corpus of 21,580 sentences annotated for relevance to the 21 BDI-II symptoms. This resource, coupled with their exploration of large language models (LLMs) like ChatGPT and GPT-4 as potential assessors, underscored the feasibility of employing LLMs in complex annotation tasks.

Guecha et al. [22] from the DS@GT team at eRisk 2024 highlighted the efficacy of sentence transform-

ers in representing text data for symptom detection. Their findings emphasized that the choice of model and feature representation significantly impacts system performance. The BDI-Sen dataset [30], also developed by PÃľrez et al., offered a sentence-level dataset aligned with BDI-II symptoms, facilitating research in symptom-specific detection. This dataset's structure and annotation schema influenced our data preparation and modeling strategies. In the realm of explainable AI, Dalal et al. [31] proposed the Knowledge-infused Neural Network (KiNN), integrating domain-specific knowledge and commonsense reasoning to enhance the interpretability of depression detection. Their approach demonstrated improved performance and provided user-level explanations, aligning with our objective of transparent symptom detection models.

Collectively, these studies inform and contextualize our methodology, which combines transformer-based architectures for multilabel classification with LLM-driven query generation [32, 33] to enhance the retrieval and identification of depressive symptoms from user-generated text.

3. Task Overview

We participated in Task 1 of eRisk 2025 Lab[1]. Essentially, the task is to rank the sentences from a collection according to their relevance to symptoms with the Beck Depression Inventory-II (BDI-II) [34]. The BDI-II consists of a set of 21 questions related to symptoms of depression such as sadness, guilty feelings, suicidal thoughts, and others, where each question indicates one of the symptoms. Each BDI-II question consists of four options based on the severity of the particular symptoms, based on the selection of the option for the question a score is calculated, which leads to identifying whether a person is depressed or not. For example, a score below 14 indicates minimal, and a score more than 29 indicates severe [35].

In this study, we focus on two points: (1) Finding relevant sentences for a specific symptom, and (2) Proposing the best strategy to derive queries from describing the BDI-II symptoms in the questionnaire. A sentence can be treated as relevant for a symptom if it is associated with any of the categories of regarding symptom, even if the user mentions they do not suffer from the given symptom.

4. Methodology

In this section, we discuss adding new terms to the existing query, which leads to query expansion, where we use semantic similarity. In particular, two methods are proposed for query expansion, that use word embeddings. On the one side, we focus on the nearest neighbour-based approach [36, 37, 38] to expand the query by adding terms that are aligned with query terms in the embedded space, as described in Section 4.1. On the other side, we utilize the Relevance model [39] to find similar terms for expanding the query set.

4.1. preRetrieval kNN method

Given a query set Q=$\{q_1, q_2, q_3, ...q_{|m|}\}$, this method first builds a set C of candidate expansion terms by finding which are semantically similar to q_i, as described in Equation 1,

$$C = NN(Q) = \bigcup_{q \in Q} NN(q) \tag{1}$$

A term t is considered as the nearest neighbor of the query Q, based on high mean cosine similarity between term t with all the terms of query Q, as presented in Equation 2.

$$\sigma(t, Q) \leftarrow \frac{1}{|Q|} \sum_{q \in Q} \xi(t, q) \tag{2}$$

[1] https://erisk.irlab.org/

```python
def generate_refined_query(symptom: str, sentences: list[str]) -> str:
    "You are a senior psychiatrist with 10 years of experience. Given the
    following sentences that describe the symptom"
    ## Instruction
     f"'{symptom}', analyze them carefully and generate a refined and
     well-structured query that effectively captures the essence of the symptom
     for better search and retrieval results. Ensure that the refined query
     maintains the meaning of the sentences while making it concise, coherent,
     and truly representative of the symptom as a whole. Your goal is to create
     a query that accurately conveys the patient's experience and distress
     in a way that enhances understanding and information retrieval."

     print("Sentences:")
     for s in sentences:
         print(s)

     print("\nRefined Query:")
     # Here, an NLP model like BERT could be used to synthesize the refined query.
     # Placeholder return statement.
     return "<refined query here>"
```

Figure 1: The structure of the prompt used in query formulation, which emulates a senior psychiatrist's approach to synthesizing a refined, concise, and coherent query from multiple symptom-descriptive sentences to enhance information retrieval and understanding.

4.2. postRetrieval kNN method

In this approach, a set of pseudo-relevant documents (PRD) is leveraged for constructing the candidate expansion terms set. In addition, it reduces the effort to search the candidate expansion terms from across the entire collection, by considering only the terms present in PRD. The remaining step of constructing the expanded query set is analogous to the previously mentioned approach, specifically, following Equation 2.

4.3. Large Language Based model: Query expansion

To further enrich the query formulation process, we leverage large language models (LLMs) to generate refined and symptom-specific queries for each of the 21 BDI-II symptoms, as shown in Figure 1. This method differs from embedding-based expansion by utilizing the generative and reasoning capabilities of LLMs to construct concise and expressive queries grounded in actual patient statements.

Given a particular symptom, we randomly select example sentences from the given dataset labeled with that symptoms. These few-shot examples are used as input to LLMs, guided by a carefully designed prompt. Multiple prompt templates were evaluated during development, and the most efficient one was chosen due to the quality and consistency of the produced queries.

Formally, let $S = \{s_1, s_2, ..., s_k\}$ denote the set of selected sentences for a given symptom. These are concatenated to form a prompt P for the LLM, which then outputs a refined query Q_{LLM}:

$$Q_{\text{LLM}} = \text{LLM}(P(S)) \tag{3}$$

The generated query Q_{LLM} is subsequently used with a classical retrieval model (BM25) to identify relevant sentences from the corpus for each symptom. This LLM-based approach enables more expressive and context-aware queries, which in turn improves the relevance of retrieved documents.

Table 1
Textual information is stored in ⟨TEXT⟩ field and doc-id is stored in ⟨DOCNO⟩.

⟨DOC⟩
⟨DOCNO⟩s_53_11_3⟨/DOCNO⟩
⟨TEXT⟩ Our deaths are inevitable, and perhaps it would be easier to be spared of the cruel way the universe seems to have planned for us to pass into the next phase of our existence, we push through, clinging onto the hope that perhaps death can come upon us while we sleep, in some peaceful manner.⟨/TEXT⟩
⟨/DOC⟩

Table 2
Overall dataset summary showing entries, queries, document IDs, and relevance distribution in the provided training dataset.

Total Entries	Unique Queries	Unique DocIDs	Total rel=1	Total rel=0
21580	21	16148	4552	17028

Table 3
Query set that has been used in the baseline method.

Q1. I do not feel sad. I feel sad much of the time....
Q2. I am not discouraged about my future. I feel more discouraged about my future .....
.....
Q21. I have not noticed any recent change in my interest in sex. I am less interested.....

5. Experimental setup

5.1. Dataset

The TREC formatted sentence-tagged dataset (as shown in Table 1) was compiled for this task and a total of 3.8 million sentences were provided as the training data. To evaluate the performance of the proposed method, we used a test dataset consisting of 10.4 million sentence sets, that was provided by the eRisk lab. Additionally, these sentences were further assessed as being relevant or not for the specific query by the proposed methods. All the information is stored in the file combination of *document numebr* (DOCNO) and *sentence information* (TEXT), as shown in Table 1. Our first step is to extract sentences and remove all the associated additional information related to the data format. Next, we remove unnecessary spaces, numerals, and punctuations from all the sentences, and finally, index all the documents using Lucene[2] or use it for training a transformer-based model.

Dataset Preprocessing

The first step in the project involves gathering data from multiple sources. These sources include several TREC files and an additional CSV file containing relevant content. The TREC files represent the unlabelled part of the dataset, containing raw textual data presented in an organized XML-like structure. Each entry is enclosed within <DOC> tags and is uniquely identified by the <DOCNO> field (e.g., `s_426_0_0`). The accompanying textual content of the user's post is nested within <TEXT> tags. Each TREC file has approximately 400 records, and the combined dataset spans over 3,000 files, corresponding to a corpus of more than 3.8 million sentences. This large and varied textual material forms the basis for detecting linguistic signals of depression.

The labeled data were offered in CSV format. This information associates individual sentences with

[2]https://lucene.apache.org/

specific symptoms of depression. Each CSV row contains four columns: `query`, `q0`, `docid`, and `rel`. The `query` column represents one of the 21 psychological symptoms listed in the Beck Depression Inventory-II (BDI-II) âĂŤ a clinically validated self-report tool used extensively to rate the severity of depression. The `docid` column maps the sentence back to its origin in the TREC files. The `rel` field indicates the relevance score âĂŤ 1 for relevant, 0 for irrelevant. The `q0` column is a placeholder and does not influence the analysis. After collecting the data, it was integrated into a unified CSV format for ease of processing and management. The final dataset had two columns: `text` and `symptom`.

Data Cleaning. To prepare the raw textual data for downstream tasks, extensive cleaning was performed. This included removing:

- emojis, hyperlinks, markdown, and HTML-like tags;
- noisy artifacts like repeated punctuation and structural remnants from forum writing;
- extra whitespaces.

These steps ensured grammatically and visually consistent sentences, facilitating accurate model training. Post-cleaning, the data distribution was analyzed. The dataset was found to be nearly balanced across the 21 classes, minimizing model bias.

5.2. preRetrieval kNN method

In Section 4.1, we illustrate that terms are selected based on the similarity score. In addition, this step involves data pre-processing, word2vec encoding, and computing similarity scores. We first pre-processed the dataset, which is then encoded using word2vec. Similarly, encoded all the terms of each query Q (BDI-II questions) using word2vec. Finally, compute the similarity between all the terms of the corpus with each query term q_i, and a term t is considered semantically similar to the query term q_i if the score is more than 0.8, which leads to constructing candidate expansion terms for each query term q_i. In the next stage, we aim to select terms to add to the query Q. In Equation 2, we show that terms are selected from the C based on the mean similarity score with all terms of query Q, and the result is presented in Table 4. This is considered as `Run 0`.

5.3. postRetrieval kNN method

In this approach, word embedding based relevance model is presented. Utilizing the hidden relevance model R, relevance documents are sampled for the query Q=$\{q_1, q_2, ...q_{|k|}\}$. Top-ranked documents are considered pseudo-relevant documents, in the absence of relevance judgment. The probability of selecting a term w from this model, denoted by $P(w|R)$, is approximated by $P(w|Q)$. The probability $P(w|Q)$ is computed by following technique: the term w is selected conditional, together with $q_1, q_2, ...q_k$ from the same distribution, underlying the top document. In Equation 4, we show the probability computation of $P(w|R)$, where the number of relevant documents is M. Eventually, the term has been selected as the possible expansion term, having the high $P(w|R)$ score.

$$P(w|R) = \sum_{D \in M} P(w|D) \prod_{q \in Q} P(w|D) \tag{4}$$

Now, the candidate expansion terms are stored in C for the query Q. Subsequently, we follow the Equation 2, for selecting term t for the expansion of query Q.

$$P'(w|R) = \mu P(w|R) + (1 - \mu) P(w|Q) \tag{5}$$

We utilized Equation 5 as well, for selecting the terms for building the candidate expansion terms for query Q. Eventually, we select top 5 and top 10 documents respectively, which is considered as `Run 2` and `Run 3`.

5.4. Language Model Based

In this study, we developed a custom multilabel text classification system based on the BERT-based transformer architecture to address the multilabel classification of psychological symptoms. This model was specifically tailored for Task 1 of the CLEF eRisk 2025 challenge, which involves identifying the presence of multiple depressive symptoms from given sentences.

We explored different transformer based model, such as `RoBERTaModel` [40], `BERT-large` [16], `Xlnet` [41] from the Hugging Face Transformers library and appended a linear classification head with a sigmoid activation function to handle multilabel output. The model uses the first token's hidden representation (analogous to [CLS]) from the final hidden state as a pooled embedding. A dropout layer with a probability of 0.1 was applied to reduce overfitting before passing the representation through the final linear layer. The model outputs logits corresponding to the presence or absence of each symptom class, with a binary cross-entropy loss computed across all labels.

Dataset Preparation and Label Encoding. The input dataset, sourced from the CLEF eRisk 2025 challenge, consisted of user sentences annotated with symptom labels. Each symptom label corresponds to a class from the Beck Depression Inventory (BDI), resulting in a total of 21 unique symptom classes. The data was first shuffled and split into training (75%), validation (10%), and testing (15%) sets.

Each label was mapped to a unique integer identifier and subsequently transformed into a one-hot encoded format, enabling multilabel classification. This approach allowed the model to predict multiple symptoms per sentence, capturing the complex nature of user-reported mental health issues.

Training Strategy. Training was conducted using the Hugging Face `Trainer` API, customized via subclassing to compute binary cross-entropy loss for multilabel classification. The model was trained for 10 epochs using the AdamW optimizer with a learning rate of 8e-5 and a weight decay of 0.01. A batch size of 8 was used for both training and evaluation to accommodate GPU memory constraints. The training pipeline included regular logging every 10 steps to monitor loss and metric evolution. Evaluation metrics included micro-averaged accuracy, precision, recall, and F1-score, reflecting the multilabel nature of the problem. Predictions were thresholded at 0.5 using the sigmoid-activated logits to derive final binary predictions for each label.

This custom RoBERTa-based multilabel classifier was successfully implemented to handle the challenging task of symptom identification from user-generated textual data. By leveraging transfer learning with a robust transformer architecture and tailoring the loss function and metrics for multilabel scenarios, the model forms a strong baseline for future work in symptom detection and early mental health risk assessment. This is considered as `Run 4`.

6. Results

In Table 4, we present the performance of our various runs submitted for Task 1. Among the five evaluated approaches, the `Run 4` (classification) run yields the best overall performance across all metrics, achieving the highest Average Precision (AP) of 0.068, R-Precision of 0.157, Precision@10 of 0.409, and NDCG of 0.228. This run outperformed the others, indicating the effectiveness of the corresponding query expansion and retrieval strategy used.

For this task, we uploaded a total of 5 runs with different configurations and thresholds for the *Search for Symptoms of Depression*.

- **Run 0**: This run consists of pre-retrieval k-nn method.
- **Run 1**: This run consists of in-context learning based few-shot learning approach.
- **Run 2**: For this run we employed post-retrieval approach where consider first 10 retrieved documents as relevant.
- **Run 3**: This run uses the same method like previous, the one difference is here we consider first 5 documents as the relevant one.

Table 4
Results of team THINKIR for Task 1 in ranking-based evaluation. Best performance in terms of the individual evaluation metrics in bold.

Run	AP	R-PREC	P@10	NDCG
0	0.003	0.010	0.000	0.030
1	0.015	0.049	0.133	0.073
2	0.064	0.148	0.400	0.213
3	0.060	0.151	0.409	0.212
4	**0.068**	**0.157**	**0.409**	**0.228**

- **Run 4**: This run consists of running a classification model obtained through the fine-tuned RoBERTa-base with the given training set.

The two runs based on `pseudo relevance feedback`, namely `Run 2` and `Run 3`, performed competitively, with AP scores of 0.064 and 0.060 respectively. These runs considered different numbers of top documents (10 and 5) for feedback, and showed similar values in P@10 (0.400 and 0.409) and NDCG (0.213 and 0.212), slightly trailing the `2025` run.

On the other hand, the `Run 1` approach yielded significantly lower performance, with an AP of 0.015 and NDCG of 0.073, indicating that the few-shot prompting strategy requires further tuning or might not generalize well for this task setting. The `Run 0` run, which used a similarity-based ranking approach, had the weakest results across all metrics, with near-zero Precision@10 and an AP of just 0.003.

7. Conclusion

In this work, we explored and evaluated multiple query expansion strategies to improve retrieval performance for Task 1 of the CLEF eRisk 2025 challenge. Our experiments included diverse approaches such as few-shot prompting, rank similarity-based retrieval, classical pseudo-relevance feedback (PRF) with varying feedback document counts, and a custom-designed retrieval run denoted as `2025`.

The experimental results demonstrate that the `2025` run, which combined effective query expansion with an optimized retrieval framework, achieved the best overall performance across all four evaluation metrics: AP, R-Precision, P@10, and NDCG. This indicates the strength of our method in capturing and retrieving relevant content in early risk detection tasks.

Our pseudo-relevance feedback variants (`pseudo relevance 10` and `pseudo relevance 5`) also performed competitively, reaffirming the effectiveness of PRF-based expansion techniques, especially in scenarios where limited labeled data is available. These methods, despite their simplicity, produced consistent and strong results close to the top-performing run.

On the contrary, the `few-shot query (Run 1)` and `Run 0` approaches lagged behind significantly, suggesting that while innovative, these strategies require further refinement or additional task-specific tuning to be viable in this domain.

Overall, our study reinforces the value of traditional IR methods like pseudo-relevance feedback, while also highlighting the potential of hybrid strategies that intelligently blend learned models and classic techniques. In future work, we aim to enhance the robustness of few-shot and ranking similarity-based methods, explore neural retrieval models, and incorporate domain-specific knowledge to further improve the retrieval quality in mental health-related search tasks.

8. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT as a writing assistant.

References

[1] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, A. Ulhaq, Depression detection from social network data using machine learning techniques, Health information science and systems 6 (2018) 1–12.

[2] A. Thapar, S. Collishaw, D. S. Pine, A. K. Thapar, Depression in adolescence, The lancet 379 (2012) 1056–1067.

[3] A. Kumar, A. Sharma, A. Arora, Anxious depression prediction in real-time social data, arXiv preprint arXiv:1903.10222 (2019).

[4] L. H. Andrade, J. Alonso, Z. Mneimneh, J. Wells, A. Al-Hamzawi, G. Borges, E. Bromet, R. Bruffaerts, G. De Girolamo, R. De Graaf, et al., Barriers to mental health treatment: results from the who world mental health surveys, Psychological medicine 44 (2014) 1303–1317.

[5] H. P. . (Group), Healthy people 2010, volume 2, US Department of Health and Human Services, Healthy People 2010, 2000.

[6] W. W. Eaton, C. Muntaner, C. Smith, A. Tien, M. Ybarra, Center for epidemiologic studies depression scale: Review and revision, The use of psychological testing for treatment planning and outcomes assessment (2004) p363–p377.

[7] K. Kroenke, R. L. Spitzer, J. B. Williams, The phq-9: validity of a brief depression severity measure, Journal of general internal medicine 16 (2001) 606–613.

[8] D. J. Dozois, K. S. Dobson, J. L. Ahnberg, A psychometric evaluation of the beck depression inventory–ii., Psychological assessment 10 (1998) 83.

[9] M. Hamilton, A rating scale for depression. journal of neurology, Neurosurgery and Psychiatry 23 (1960) 56–62.

[10] G. Rao, Y. Zhang, L. Zhang, Q. Cong, Z. Feng, Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums, IEEE Access 8 (2020) 32395–32403.

[11] R. Skaik, D. Inkpen, Using twitter social media for depression detection in the canadian population, in: Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference, 2020, pp. 109–114.

[12] A. S. Uban, B. Chulvi, P. Rosso, Multi-aspect transfer learning for detecting low resource mental disorders on social media, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3202–3219.

[13] K. A. Hambarde, H. Proenca, Information retrieval: recent advances and beyond, IEEE Access (2023).

[14] 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.

[15] 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.

[16] J. Devlin, M.-W. Chang, K. Lee, K. 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, volume 1 (long and short papers), 2019, pp. 4171–4186.

[17] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, et al., A survey on in-context learning, arXiv preprint arXiv:2301.00234 (2022).

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[19] C. Qin, A. Zhang, C. Chen, A. Dagar, W. Ye, In-context learning with iterative demonstration

selection, arXiv preprint arXiv:2310.09881 (2023).
[20] A. Barachanou, F. Tsalakanidou, S. Papadopoulos, Rebecca at erisk 2024: Search for symptoms of depression using sentence embeddings and prompt-based filtering, in: CLEF eRisk, 2024.
[21] B. H. Ang, S. D. Gollapalli, S.-K. Ng, Nus-ids@erisk2024: Ranking sentences for depression symptoms using early maladaptive schemas and ensembles, in: CLEF eRisk, 2024.
[22] D. Guecha, A. Potdar, A. Miyaguchi, Ds@gt erisk 2024: Sentence transformers for social media risk assessment, in: CLEF eRisk, 2024.
[23] A. M. MÃąrmol-Romero, A. Moreno-MuÃśoz, P. ÃĄlvarez Ojeda, et al., Sinai at erisk@ clef 2024: Approaching the search for symptoms of depression and early detection of anorexia signs using natural language processing, in: CLEF eRisk, 2024.
[24] R.-M. Hanciu, Mindwaveml at erisk 2024: Identifying depression symptoms in reddit users, in: CLEF eRisk, 2024.
[25] A. Pardo BascuÃśana, I. Segura Bedmar, Apb-uc3m at erisk 2024: Natural language processing and deep learning for the early detection of mental disorders, in: CLEF eRisk, 2024.
[26] H. Oliveira, et al., Bioinfo@uavr at erisk 2020: Early risk detection of depression, in: CLEF, 2020.
[27] C. Spartalis, G. Drosatos, A. Arampatzis, Transfer learning for automated responses to the bdi questionnaire., in: CLEF (Working Notes), 2021, pp. 1046–1058.
[28] D. E. Losada, F. Crestani, J. Parapar, et al., Overview of erisk at clef 2024: Early risk prediction on the internet (task 1), in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[29] Ã. Pérez, et al., Depresym: A depression symptom annotated corpus and the role of llms as assessors of psychological markers, arXiv preprint arXiv:2308.10758 (2023).
[30] Ã. Pérez, et al., Bdi-sen: A sentence dataset for clinical symptoms of depression, in: Proceedings of SIGIR, 2023.
[31] S. Dalal, et al., Deep knowledge-infusion for explainable depression detection, arXiv preprint arXiv:2409.02122 (2024).
[32] S. Adhikary, P. Sen, D. Roy, K. Ghosh, A case study for automated attribute extraction from legal documents using large language models, Artificial Intelligence and Law (2024) 1–22.
[33] S. Sar, S. Adhikary, D. Roy, Baaf: A framework for media bias detection, in: European Conference on Information Retrieval, Springer, 2025, pp. 255–264.
[34] B. D. INVENTORY-II, Beck depression inventory-ii, The Corsini Encyclopedia of Psychology, Volume 1 1 (2010) 210.
[35] C. B. Loh, Y. H. Chan, Psychological symptoms in people presenting for weight management, Annals Academy of Medicine Singapore 39 (2010) 778.
[36] D. Roy, M. Mitra, P. Mayr, A. Chowdhury, Local or global? A comparative study on applications of embedding models for information retrieval, in: G. Dasgupta, Y. Simmhan, B. V. Srinivasan, S. Bhowmick, A. Singhee, M. Ramanath, N. Batra, A. S. Prasad (Eds.), CODS-COMAD 2022: 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD), Bangalore, India, January 8 - 10, 2022, ACM, 2022, pp. 115–119. URL: https://doi.org/10.1145/3493700.3493701. doi:10.1145/3493700.3493701.
[37] D. Roy, Word embedding based approaches for information retrieval, in: Seventh BCS-IRSG Symposium on Future Directions in Information Access, BCS Learning & Development, 2017.
[38] D. Roy, D. Paul, M. Mitra, U. Garain, Using word embeddings for automatic query expansion, arXiv preprint arXiv:1606.07608 (2016).
[39] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at trec 2004: Novelty and hard, Computer Science Department Faculty Publication Series (2004) 189.
[40] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).